DNA
Volume 6, Number 5, 1987
Mary Ann Liebert, Inc., Publishers
Pp. 493-495

SHORT COMMUNICATION

Database Bias and the Identification of Protein Coding Sequences

MICHAEL E. MOODY* and BRIAN FRISTENSKY†,‡

ABSTRACT

A simple quantitative test for the probability that an open reading frame actually codes for a protein has been described by Tramontano and Macchiato (1986). However, their test is only valid for the special case in which both coding and noncoding sequences are represented equally. We present a generalized adaptation of their method that uses estimates for the relative proportions of coding and noncoding sequences to provide a more accurate prediction.

A COMMONLY ENCOUNTERED PROBLEM in hypothesis testing is to distinguish between two alternatives based upon some measurement for which the alternatives differ only in their *distribution* of the quantity measured. For example, measurement of α-fetoprotein level is used to distinguish between normal fetuses and those with certain neuropathies; normal and abnormal fetuses (as groups) differ only in the distribution of the titer of this protein. As another example, consider a patient who tests positive for the presence of some disease: it is desired to know the probability, given the test result, that the patient is actually afflicted. Unless the test is a perfect predictor of the disease, the conditional probability will depend upon the incidence of the disease, since the conditional chance of a false positive result will be greater when the disease is at an ebb than when it is epidemic. In general, the relative occurrence of the two alternatives in the population as a whole will influence predictions based upon such measurements.

Tramontano and Macchiato (1986) have recently defined the "information value" of an open reading frame (ORF), which they use to classify DNA sequences by a probabilistic criterion, as either coding, or noncoding. Simply put, the information value of a given codon is a measure of the degree to which mutation of that codon would likely result in a change in hydrophobicity (Tramontano and Macchiato, 1982, 1986). They observed that the distribution of the information value for coding sequences is approximately normal, with a mean of 2.25 and a standard deviation of .15; the corresponding values for noncoding sequences are 2.5 and .25. The differences between coding and noncoding are significant. However, their analysis failed to account properly for the relative distribution of coding *vs.* noncoding sequences. We present below the generalized expressions, with an analysis of their applicability.

Suppose the information value for a coding sequence (V_c) has an underlying normal (Gaussian) probability density, with mean μ_c and standard deviation σ_c; the information value for noncoding sequences ($V_{n/c}$) is normal, with mean μ_{nc} and standard deviation σ_{nc}. The respective distributions functions are therefore

$$P[V_c \leq v] = \frac{1}{\sqrt{2\pi\,\sigma_c}} \int_{-\infty}^{v} \exp(-[(x - \mu_c)/\sigma_c]^2/2)\,dx$$

$$P[V_{nc} \leq v] = \frac{1}{\sqrt{2\pi\,\sigma_{nc}}} \int_{-\infty}^{v} \exp(-[(x - \mu_{nc})/\sigma_{nc}]^2/2)\,dx$$

To complete our specification, we posit that the "genome" is an independent collection of coding and noncoding sequences, so that the probability that a randomly selected sequence is noncoding is some (usually unknown) value, θ. Thus, if V represents the information value of a randomly selected sequence, it is manifestly normally distributed with mean $\mu = (1 - \theta)\mu_c + \theta\mu_{nc}$ and standard deviation $\sqrt{((1-\theta)\sigma_c)^2+(\theta\sigma_{nc})^2}$. If we are presented with a sequence with information value v, we compute the probability that the sequence is noncoding to be

*Department of Pure and Applied Mathematics and Program in Genetics and Cell Biology and †Program in Genetics and Cell Biology, Washington State University, Pullman, WA 99164-2930.
‡Present address: Department of Botany, North Carolina State University, Raleigh, NC 27695-7612.

$$P_{nc} = \frac{\left\{\theta \exp(-[(v - \mu_{nc})/\sigma_{nc}]^2/2)\right\}/\sigma_{nc}}{\left\{(1 - \theta)\exp(-[(v - \mu_c)/\sigma_c]^2/2)\right\}/\sigma_c + \left\{\theta \exp(-[(v - \mu_{nc})/\sigma_{nc}]^2/2)\right\}/\sigma_{nc}} \qquad (1)$$

where $P_{nc} = P[nc \mid V = v]$. The explicit inclusion of the weight θ differentiates our treatment from that of Tramontano and Macchiato; their expression is correct (apart from a missing factor of ½ in the exponents of the numerator and the first term of the denominator, and the absent power of 2 in the exponent of the second term of the denominator) only in the special case of an equally mixed distribution of coding and noncoding sequences, θ = ½.

Our modification, though technically simple, is of consequence. As θ ranges from 0 (all sequences are coding) to 1 (all sequences are noncoding), the assessment of the chance that the sequence is noncoding, given any information value *v must* increase from 0 to 1; this is reflected in our formulation, but not that of Tramontano and Macchiato. Since θ is generally unknown, this diminishes the utility of their method and can lead to seriously incorrect estimates of their method and can lead to seriously incorrect estimates if ignored altogether. However, since the overall distribution of information value (V) is normal according to our assumptions, if we have an estimate of the mean μ we may rearrange it to obtain the following estimate of θ:

$$\theta = \frac{\hat{\mu} - \hat{\mu}_c}{\hat{\mu}_{nc} - \hat{\mu}_c} \qquad (2)$$

where the variables with a caret (^) are the respective estimates. This is also the maximum likelihood estimate for the mixing parameter. It is relatively straightforward to show that

$$\theta = \frac{n_{nc}}{n_{nc} + n_c} \qquad (3)$$

where n_{nc} and n_c are, respectively, the number of noncoding and coding sequences in a sample. We remark that this would be the logical estimator for P_{nc} in the absence of any knowledge of the information value of the sequence. To use this method to evaluate P_{nc} for a given sequence with information value v, it is first necessary to estimate the means and standard deviations μ_{nc}, μ_c, σ_{nc}, and σ_c from an appropriate sequence database, estimate θ by employing equation (3) [or (2)], and finally substitute into (1) and evaluate the resulting expression. The dependence of P_{nc} on v for various values of θ is shown in Fig. 1, which uses the means and standard deviations given by Tramontano and Macchiato, quoted above.

Since databases are inherently biased collections of sequences (e.g., coding sequences are disproportionately represented), it is useful to constrain the sequences to some subset of a sequence database for which we can estimate θ, the means μ_c and μ_{nc}, and the standard deviations σ_c and σ_{nc}. For example, if the sequence (with unknown coding properties) is known to be of mitochondrial origin, one would calculate the estimates only for mitochondrial sequences. To minimize the error caused by the overlapping information value distributions of coding and noncoding sequences, these parameters can be estimated using ORFs whose lengths are no shorter than that of the unknown sequence. Although the interpretation of these estimates is complicated by the nonrandom character of sequence databases, neglecting θ entirely can lead to serious errors in the assignment of sequence identity (see Fig. 1).

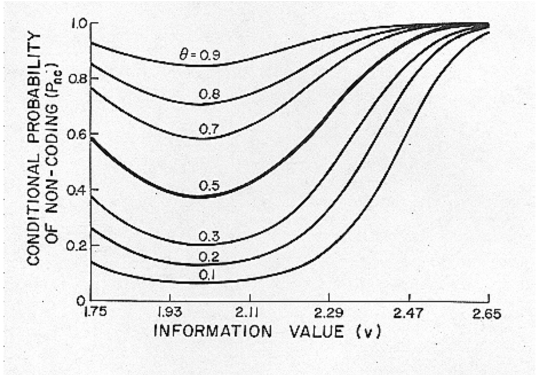

FIG. 1. Plot of P versus information value v, for various values of θ. The curves correspond, bottom to top, to θ = .1, .2, .3, .5, .7, .8, and .9. Also, μ_c = 2.25, μ_{nc} = 1.5, σ_c = .15, and σ_{nc} = .5. The curve for θ = .5 is highlighted, as this corresponds to Fig. 1d of Tramontano and Macchiato (1986).

The predictions based on information value can be compared to the TESTCODE algorithm of Fickett (1982). TESTCODE provides a prediction of coding, noncoding, or no opinion based on the degree of bias detected in nucleotide usage within codons. TESTCODE has proven reliable in classifying sequences for two reasons. First, TESTCODE takes into account the relative distributions of eight parameters between coding and noncoding sequences in an actual database. Second, Fickett has empirically determined that TESTCODE's acuracy is optimal when used with ORFs longer than 200 bp. Since the decision criteria used by TESTCODE are fundamentally different from those used herein, our corrected form of the information value function should provide a useful complement to TESTCODE.

An important conclusion to be drawn from this exercise is that any method for prediction of the functional nature of an unknown sequence will be subject to additional error if it does not take into account the relative proportions of the different types of sequences in the population. For example, there are several methods for identifying splice junctions in DNA sequences (Nakata *et al*., 1985; Iida, 1985; Staden, 1984; Harr *et al*., 1983), but in general these methods find more false positive identifications than actual splice junctions, owing to the fact that chance will create many sites that are similar to true junctions.

Although the effort required to estimate the necessary population parameters for our analysis can be substantial, we feel that the resultant increased accuracy of the predictive test more than justifies such effort.

REFERENCES

FICKETT, J.W. (1982). Statistical evaluation of the coding capacity of complementary DNA strands. Nucleic Acids Res. **10**, 5303-5318.

HARR, R., HÄGGSTRÖM, M., and GUSTAFSON, P. (1983). Search algorithm for pattern match analysis of nucleic acid sequences. Nucleic Acids Res. **11**, 2943-2957.

IIDA, Y. (1985). Splice-site signals of mRNA precursors as revealed by computer search. J. Biochem. **12**, 1173-1179.

MACCHIATO, M.F., and TRAMONTANO, A. (1982). Thermodynamic approach to a possible theory of the evolution of a genetic code. Z. Naturforsch. **37c**, 1031-1037.

NAKATA, K., KANEHISA, M., and DELISI, C. (1985). Prediction of spliced junctions in mRNA sequences. Nucleic Acids Res. **13**, 5327-5340.

STADEN, R. (1984). Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. **12**, 505-519.

TRAMONTANO, A., and MACCHIATO, M.F. (1986). Probability of coding of a DNA sequence: An algorithm to predict translated reading frames from their thermodynamic characteristics. Nucleic Acids Res. **14**, 127-135.

Address reprint requests to:
Dr. Michael E. Moody
Department of Pure and Applied Mathematics
Washington State University
Pullman, WA 99164-2930

Received for publication February 23, 1987, and in revised form June 29, 1987.