Volume 14 Number 1 1986 Nucleic Acids Research

A statistical method for correlating tRNA sequence with amino acid specificity

Taskin Atilgan, Hugh B.Nicholas, Jr. and William H.McClain

Department of Bacteriology, University of Wisconsin-Madison, Madison, WI 53706, USA

Received 15 July 1985

ABSTRACT

A statistical method for finding the nucleotide positions in tRNA sequences that correlate with amino acid specificity has been developed. The procedure involves finding the subset of nucleotide positions and groups of positions where the marginal density of one amino acid tRNA class does not overlap that of any other amino acid class. The procedure is an application of a statistical method known as the Expectation Maximization algorithm.

INTRODUCTION

We are developing computer-assisted methods to search tRNA sequences for nucleotide positions that correlate with amino acid specificity. Our goal is to obtain predictive information for laboratory experiments designed to disclose the nucleotides in tRNA molecules that carry the amino acid specificity determinants for the aminoacyl-tRNA synthetases. The method described below makes use of a data set containing a number of isoacceptor tRNA chains for each amino acid. The method locates the nucleotide positions and combinations of positions unique to each amino acid class. This paper presents a statistical formulation of the problem, followed by development of an algorithm[1] to obtain a solution. In the algorithm, the subset of nucleotide positions is found over which the density of one amino acid tRNA does not overlap the density of any other amino acid class. The density represents a multivariate histogram of four cells at each variable position in a tRNA sequence.

RESULTS

Statistical Formulation - Let $\underline{X}_i$ $i=1,2,...,N$ denote the vectors, one for each tRNA sequence. The number of dimensions in each vector corresponds to

[1] A Fortran 77 listing of the algorithm will be provided free of charge on written request to William H. McClain.

the number of residues in the tRNA chain (e.g. 76). Now, let $\{(\underline{X}_1, Y_1), \ldots(\underline{X}_N, Y_N)\}$ be the data set complete with their identifiers, Y_i, $i=1,\ldots,N$. The Y_i's are integers taking values from 1 to M, identifying the class to which the observation (tRNA sequence) $\underline{X}_i$ belongs.

Knowledge of the complete sequence vector $\underline{X}_i$ gives us the value of the identifier Y_i. The question is: are there subsets of variables (positions and nucleotides in those positions) in vector $\underline{X}$ which by themselves are sufficient to assign a tRNA sequence to an amino acid class? The answer will allow us to precisely identify the tRNA positions that correlate with the 20 amino acid classes.

Let $\underline{X}_{(k_1,\ldots,k_e)} = (X_{k_1},\ldots,X_{k_e})$ be the subset of variables constructed by taking $k_1{}^{th},\ldots,k_e{}^{th}$ variables of vector $\underline{X}$. Consider the conditional probability (reference 1) that the identifier of i^{th} observation $\underline{X}_i$ is $Y_i=j$, given that we have the subset $\underline{X}_{i(k_1,\ldots,k_e)}$:

$$[1] \quad P(Y_i=j|X_{i(k_1,\ldots,k_e)}) = \frac{\theta_{j(k_1,\ldots,k_e)} f_j(X_{i(k_1,\ldots,k_e)})}{\sum\limits_{h=1}^{M} \theta_{h(k_1,\ldots,k_e)} f_h(X_{i(k_1,\ldots,k_e)})}$$

$$= \Pi_{i(k_1,\ldots,k_e)}(j)$$

$j=1,\ldots,M$ (number of groups); $i=1,\ldots,N$ (number of observations)
where $\theta_{h(k_1\ldots,k_e)}$ are the marginal prior probabilities and $f_h(\underline{X}_{i(k_1,\ldots,k_e)})$ are marginal densities of M classes ($h=1,2,\ldots,M$). Marginal densities $f_h(X_{(k_1,\ldots,k_e)})$ are e-variate histograms, constructed by using the observations, $\underline{X}_i$'s, belonging to the class h.

Suppose we are equally likely to assign a tRNA sequence to any amino acid class if we consider only a subset of the variables. Then we initially take $\theta_{h(k_1,\ldots,k_e)} = \frac{1}{M}$ for $h=1,2,\ldots,M$.

We have

$$\Pi_{i(k_1,\ldots,k_e)}(j) = \frac{f_j(X_{i(k_1,\ldots,k_e)})}{\sum\limits_{h=1}^{M} f_h(X_{i(k_1,\ldots,k_e)})}$$

$$j=1,\ldots,M;\ i=1,\ldots,n_h;\ \sum_{h=1}^{M} n_h=N.$$

Here, we are <u>allocating</u> n_h observations of class h to M classes by fractions, using only the information coming from the subset $\underline{X}_{(k_1,\ldots,k_e)}$ of observation vectors.

Next, we aggregate these fractions over the observations coming from a specific class to obtain the current value $\theta^{(p+1)}_{h(k_1,\dots,k_e)}$:

$$[2] \qquad \theta^{(p+1)}_{h(k_1,\dots,k_e)} = \frac{1}{n_h} \sum_{i=1}^{n_h} \Pi_{i(k_1,\dots,k_e)}(h),$$

where n_h is the number of sequences belonging to class h (=1,2,...,M). Then, we use the new $\theta^{(p+1)}_{h(k_1,\dots,k_e)}$ values in the "allocation" step to calculate new values for $\Pi_{i(k_1,\dots,k_e)}(j)$, and continue the iteration until convergence. The final value of $\theta_{h(k_1,\dots,k_e)}$, obtained at the end of iteration, can be taken as a measure of identifying power of the subset $\underline{X}_{(k_1,\dots,k_e)}$ for category h. This statistical algorithm is analogous to the EM algorithm of Dempster et al. (reference 2), with the "allocation" step corresponding to the E step (Expectation step) and the "aggregation" step corresponding to the M step (Maximization step). To select subsets that are identifiers we consider:

$$[3] \qquad \max_{(k_1,\dots,k_e)\varepsilon A} \theta_{h(k_1,\dots,k_e)},$$

where A is the set of all possible combinations of variables with $k_1<k_2<\dots<k_e$, e=1,2,...,L, and L is 76 (or more), and $0 < \theta_{h(k_1,\dots,k_e)} <1$. If $\theta_{h(k_1,\dots,k_e)} = 1$ for class h when the subset $\underline{X}_{(k_1,\dots,k_e)}$ is used, then the subset $\underline{X}_{(k_1,\dots,k_e)}$ is a perfect identifier for class h as far as the given set of observations $\underline{X}_1,\dots,\underline{X}_N$ is concerned. Therefore, first we try to ascertain if there are subsets over which

$$\theta_{j(k_1,\dots,k_e)} = \begin{cases} 1 \text{ for } j=h \\ 0 \text{ for } j\neq h \end{cases} \quad j=1,2,\dots,M.$$

From equation [1] we see that $\theta_{h(k_1,\dots,k_e)} = 1$ is achieved when the density $f_h\left(\underline{X}_{i(k_1,\dots,k_e)}\right)$ is non-overlapping with the density $f_j\left(\underline{X}_{i(k_1,\dots,k_e)}\right)$ for all $j\neq h$, j=1,..., M. That is,

$$\int f_h\left(X_{(k_1,\dots,k_e)}\right) f_j\left(X_{(k_1,\dots,k_e)}\right) dX_{(k_1,\dots,k_e)} = 0 \text{ for } h\neq j.$$

Computer Algorithm To Find Non-overlapping Subsets - From equations [1] and [2] we obtain

$$\theta_{h(k_1,\dots,k_e)} = 1$$

if

$$f_h\left(X_{i(k_1,\dots,k_e)}\right) \begin{cases} \neq 0 & \text{for } i = 1,\dots,n_h \\ = 0 & \text{otherwise} \end{cases}$$

a)

Amino acid	Position number Nucleotide 1	2	3	4	5
ALA-1	G	G	G	G	G
ALA-2	G	G	G	G	C
ARG-1	G	C	A	U	C
ARG-2	G	C	A	U	C
ARG-3	G	C	G	C	C
LEU-1	G	C	G	A	A
LEU-2	G	C	C	G	G
LEU-3	G	C	C	C	G

b)

	ALA-1 1	2	3	4	5	ALA-2 1	2	3	4	5
ALA-1	G	G	G	G	G					
ALA-2						G	G	G	G	C
ARG-1	S	D	D	D	D	S	D	D	D	S
ARG-2	S	D	D	D	D	S	D	D	D	S
ARG-3	S	D	S	D	D	S	D	S	D	S
LEU-1	S	D	S	D	D	S	D	S	D	D
LEU-2	S	D	D	S	S	S	D	D	S	D
LEU-3	S	D	D	D	S	S	D	D	D	D

		ALA-1	ALA-2	
One ALA sequence	One-position discriminators	{2}	{2}	Allocation

	Two-position discriminators	{(3,4)}
		{(3,5)}
			{(3,4)}
			{(4,5)}	

All ALA sequences	One-position discriminators	{2}		Aggregation
	Two-position discriminators	{(3,4)}		

Figure 1. Identification of Discriminators.
(a) The first five nucleotide residues of two alanine tRNAs, three arginine tRNAs, and two leucine tRNAs. The sequences are arbitrarily labeled (-1, -2, -3). The sequence LEU-2 is artificial (to help illustrate the method).
(b) Discrimination matricies for the two ALA sequences. D, different nucleotide. S, same nucleotide. Allocation and aggregation steps are indicated at the bottom right. Nucleotide positions in the intersection-sets are given in the last two rows.

when the e-variate histogram of class h over the subset $\underline{X}_{(k_1,\ldots,k_e)}$ does not overlap with that of any other amino acid class.

To obtain the non-overlapping subsets, the positions in individual sequences of one amino acid class are compared with the same positions in individual sequences belonging to the other 19 amino acid classes. As a result of these comparisons, we obtain n_h matrices of dimension $(N-n_h)xL$, where L is, again, the length of the sequence. Elements of these "discrimination matrices" are: (D), when the nucleotide of a sequence at a certain position is different from the same position of another sequence; and (S), when the nucleotide is the same. Consider the discrimination matrix for alanine tRNA-1 (ALA-1) shown in Figure 1. The uninterrupted column of D's in position 2 shows that this position discriminates ALA-1 from sequences belonging to all other amino acid classes. Position 2 is thus a discriminating position for sequence ALA-1; note that it is also a discriminating position for sequence alanine tRNA-2 (ALA-2). Pairs of positions can also serve as discriminators for a given sequence. ALA-1 is discriminated from the three arginine tRNAs (ARG-1-ARG-3) and three leucine tRNAs (LEU-1-LEU-3) sequences by the presence of D's in column 2 or 4; positions 3 & 5 are also a discriminating pair for ALA-1. Analogously, ALA-2 is discriminated by the pairs of positions 3 & 4 and 4 & 5. Locating the discriminating positions (single or multiple) for individual sequences (e.g.

ALA-1) constitutes the allocation step of our application of the EM algorithm where $\Pi_{i(k_1,\ldots,k_e)}(h) = 1$ for sequence i of class h.

In the subsequent aggregation step we find the discriminating positions that are common to all isoacceptor sequences of a given amino acid class. This operation produces the intersection of the non-overlapping subsets obtained for sequences $i=1,\ldots,n_h$ of class h. The aggregation step requires $\Pi_{i(k_1,\ldots,k_e)}(h) = 1$ for $i=1,\ldots,n_h$ to obtain $\theta_{h(k_1,\ldots,k_e)}=1$; thus, the subset $\underline{X}_{(k_1,\ldots,k_e)}$ is in the intersection-set. Figure 1b (bottom) gives the elements of the intersection-set for the indicated sequences and discrimination matrices. Position 2 is in the intersection-set of the one-position discriminators. The pair of positions 3 & 4 is in the intersection-set of the two-position discriminators. While the pairs 3 & 5 for ALA-1 and 4 & 5 for ALA-2 discriminate individual ALA sequences, they are not in the intersection-set. There are no other multiple-position discriminators (three or larger) for the ALA sequences in Figure 1. Though position 2 could combine with any other (or more) position to give a unique pair (or more), such combinations are redundant and thus are ignored when the algorithm is used.

DISCUSSION

The goal of this work is to develop methods that provide insight into and understanding of the structure of tRNA sequences. What makes a tRNA sequence interesting as statistical entity is its high specificity and complexity, including:

- high dimensionality -- 76 positions;
- a mixture of various tRNA types--20 amino acid classes;
- nonhomogeneity--different relationships hold between variables (positions) in different parts of the measurement vector (tRNA sequence).

A difficulty with dimensionality is that, as it increases, the data points become more sparse and spread apart. For example, a histogram that has 4 intervals (as is the case with tRNAs) in each dimension produces 4^L cells in L dimension (e.g. $L \cong 76$ with tRNA sequences). For even moderate values of L, a very large data set is needed to obtain a meaningful (i.e., predictive) histogram.

Some of the features that make tRNAs attractive as statistical entities have undoubtedly hindered identification of the amino acid information of these molecules. Traditional biochemical techniques augmented with appropriate statistical methods offer a new approach. The method presented above brings forward salient features of the data, discards the variables that mask certain aspects via the 'noise' they contribute, and provides for the analyst informative summaries of that information. It is important to emphasize that, in practice, the method described performs best with large data sets containing a number of isoacceptor tRNAs for each amino acid; this produces variation on nucleotide positions needed to reduce the size of the intersection-set.

We have applied the algorithm to a set of 65 tRNAs that function in E. coli and S. typhimurium (unpublished). Five amino acid classes had one-position discriminators; these can be identified by visual inspection of aligned tRNA sequences. Use of the computer algorithm to locate the two-position discriminators was important, however, with about ten million comparisons needed to obtain this solution. Nineteen amino acid classes had two-position discriminators; all twenty had three-position discriminators. Operating on a Digital PDP-11/23 + computer, the altorithm requires 14 min cpu time to locate nucleotide positions and combinations of positions unique to each amino acid class. It will be important to assess the predictive value of these computational results in laboratory experiments.

ACKNOWLEDGEMENTS

This research was supported by grant AI10257 from the National Institutes of Health, and by the College of Agricultural and Life Sciences and the Graduate School, University of Wisconsin-Madison.

REFERENCES

(1) Bayes, T. (1763) Philosophical Transactions of the Royal Society 53: 370-418. Reprinted (1958) Biometrika 45: 293-315.

(2) Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) J. Royal Statistical Soc. 39: 1-38.