NONLINEAR CLASSIFICATION
OF
TRANSLATIONAL INITIATION
SITES

by

William W. Brown
Department of Computer Science
University of Colorado
Boulder, Colo. 80309

CU-CS-211-81 October 1981

Completed August 1981 as CS-950 Graduate Independent Study for the M.S. degree in computer science.

Abstract: The use of nonlinear methods in the classification of mRNA gene initiation sites has two important aspects: (1) the realization of weight vector W separating gene and nongene sites not linearly separable, and (2) insights into combinations of mRNA nucleotides which are important in the start site selection process. The present paper addresses each of these areas and gives results obtained from experiments in nonlinear classification of translational initiation sites.

1.0 Introduction

In February, 1981, the author used a nonlinear application of F. Rosenblatt's Perceptron [1] to separate won and lost positions from a "dot" game called SIM. The work produced three important results: (1) that very large sets of won and lost signals could be separated by using higher order training algorithms, (2) significant generalization abilities were "learned" leading to correct placement of a high percentage of new (not in the original training set), positions, and (3) that additional heuristic information was gained by examining the weight coefficients of the higher order terms in the nonlinear space. It is believed that these results also apply to the recognition of separable patterns not of game space origin.

The present paper will examine some of these results with respect to a well known problem in molecular biology. In Sections 2 and 3 a statement of this problem is given and some background information provided. Section 4 will examine the computational framework for nonlinear methods in Perceptron learning. Finally, Section 5 will present some experimental results.

2.0 Statement of the Problem

In the field of genetic biology, considerable effort has been applied to how protein is synthesized in the cell and the relation this process has with genetic material in DNA. In particular, there is interest in the determination of translational initiation sites on strands of messenger RNA (mRNA) where the protein is built during an interaction with cellular ribosomes. These mRNA sites are thought to possess certain genetic signals that are somehow recognized by the ribosome before the protein initiation is begun (hence translational initiation site). The genetic information is contained in linear strings of mRNA units called nucleotides. A ribosome begins at a specific site and then makes a protein (a string of amino acids) specified by the sequence of nucleotides, which are read three at a time (in "codons"). The problem is to determine what arrangement of nucleotides specifies the initiation sites. (The preceding taken from discussions with G. Stormo.)

Nucleotides are of four types, called bases: thymidine, cytosine, adenine, and guanine designated T, C, A, and G, respectively. A specific arrangement of these bases constitutes a message sent (via mRNA) from the genetic DNA to the ribosome designating the particular region of mRNA as a translational initiation site [2]. Estimates have been given that the selection of an initiation site is made on the basis of 35 or 40 mRNA nucleotides. In particular, a centrally located ATG codon (or less commonly GTG) has been shown to be an almost necessary condition for a gene initiation site. The centrality has lead to a convention that the ATG codon occupy the 0-2 elements of a mRNA site having description of, say, (-30,+20) which is 51 nucleotides [3].

Using the same frame of reference, Shine and Dalgarno found that the ATG codon was preceded by filler and all or part of the following string:

TAAGGAGGT

This leads to an idealized gene initiation site description:

Filler	Shine and Dalgarno Sequence	3-9 Bases Filler	0-2	Filler	End Codon
	TAAGGAGGT		ATG		TAA

where TAA is one of three end codons.

Since many non-initiation sites may have similar Shine and Dalgarno sequences, in addition to the central ATG codon, greater insights into the interaction of bases are needed before the problem of translational initiation is solved.

3.0 Background

The first study using a Perceptron in gene initiation site determination was carried out by G. Stormo at the University of Colorado, Boulder [3]. Stormo and collegues built a data base of known gene and nongene sites (or more precisely, ribosome binding sites and not), and a language compiler ("DELILA") to aid in accessing specified strands. With this system, Stormo used a linear model of the Perceptron against 124 known gene initiation sites and increasingly large numbers of nongene sites. His procedure was to realize linear separation between the 124 gene sites and a small set of nongene sites and then to "check" the separating hyperplane against the remaining 75000+ nongene sites in the data base. Nongene sequences that were incorrectly classified were added to the training set and the process repeated until complete linear separation was achieved.

For longer strands of mRNA (101 bases), the convergence upon a linear separation functional was routine. This is not too surprising with training signals of such high dimension. In general, the necessary convexity requirements for linear separation are more likely to be satisfied when the dimension is great. As the mRNA strand lengths were shortened, Stormo found linear separation successively more difficult to achieve. Finally, a training set comprised of 51 base strands (-30,+20) yielded no separating functional after a very large number of training passes.

At first, this may seem to contradict the notion that ribosome binding site selection is based on 40 or 50 nucleotides. However, the linear model is very limited and takes no consideration of interactions between the individual bases in a given strand. Since ribosomes probably consider such higher order relationships in making the selection, it is not surprising that the 51 base strands were not linearly separable.

This lead to the present work: to use higher order applications of the linear Perceptron to achieve separation of successively shorter gene and nongene sites. By doing this, additional information concerning the importance of combinations of nucleotides is provided. As will be seen, such combinations as those in the Shine and Dalgarno sequence are very important in gene initiation site classification.

4.0 Nonlinear Methods in Perceptron Learning

The linear Perceptron model is an iterative procedure where a vector W is sought separating two sets A and B such that for signal X_n:

$$W_n = \begin{cases} W_{n-1} + X_n & \text{if } (W_{n-1}, X_n) \leq 0 \text{ and } X_n \text{ in } A \\ W_{n-1} & \text{if } (W_{n-1}, X_n) > 0 \text{ and } X_n \text{ in } A \\ W_{n-1} - X_n & \text{if } (W_{n-1}, X_n) \geq 0 \text{ and } X_n \text{ in } B \\ W_{n-1} & \text{if } (W_{n-1}, X_n) < 0 \text{ and } X_n \text{ in } B \end{cases}$$

where (W, X) is notation for the scaler product of weight vector W and training set signal X. It has been shown that this "error correction" procedure will converge upon W in a finite number of passes through the training set provided the two sets to be separated are suitably convex [4].

The usual method of applying this algorithm is to store only the indices of ones (1's) for each of many binary patterns in the training set. In this manner, a scaler product between W and any signal may be taken merely by summing the elements of W corresponding to the nonzero elements in the training signal. The error correction that may follow is done in a similar manner. This simple technique becomes critical when attempting to store huge patterns resulting from nonlinear applications.

The nonlinear algorithm itself is just an extension of the linear procedure. Instead of training on simple binary representations of training signals, an extended binary representation is employed reflecting combinations of elements in the linear signal. For example, second order separation might employ the mapping:

$$(x1, x2, \ldots, xn) \rightarrow$$

$$(x1x1, x1x2, \ldots, x1xn, x2x2, \ldots, x2xn, \ldots, xnxn)$$

where each xi is the ith element of the linear representation. From this transformation we have a new pattern containing the original linear terms (e.g. xixi, i=1,n) as well as 2nd order terms of all combinations of elements taken two at a time. The nonlinear algorithm is just the normal error correction procedure applied against the transformed signals.

From a resource standpoint, it is difficult to first transform the entire training set and then to store and operate on the expanded signals. In the second order example, the transformed signals have dimension increasing approximately by the square of the linear dimension (D) over two (D**2/2). Higher order dimensions increase even faster (at least by D**n/n! for n the order of separation required). In one experiment, discussed in Section 5.2, the training signals were each mapped into a space having a dimension exceeding 100,000!

Even with machines having virtual memory capabilities, patterns of such enormous size will soon exhaust available resources. The alternative is to store the expanded patterns on disk which is then accessed with each successive pass through the training set. This unfortunately leaves the training program "IO Bound" and is therefore not reasonable for separations requiring a large amount of training.

The solution to this problem is to store only the linear patterns (more precisely, the location of "ones" in the linear patterns), and to use a mapping to calculate the location of elements in the transformed space. The execution of this mapping for each signal during each pass exercises the host computer computationally and therefore does not involve excessive disk IO.

The nth order mapping for linear patterns of dimension D is a function Fn(c1,c2,...,cn,D) where C=(c1,c2,...,cn) is a matrix of indices defining valid combinations in the transformed space and c1<=c2<=...<=cn. The transformed index (Z) of any such combination is:

$$Z = Fn(C,D) = \sum_{i=1}^{n-1} \left[\sum_{s=1}^{D-1} Fi(Si,s) - \sum_{s=1}^{D-c(n-i)} Fi(Si,s) \right] + cn \quad (n>1)$$

$$F1(C,D) = F1(c1,D) = c1$$

where Si is a matrix of dimension i ($1 <= i <= n-1$) with all elements a dummy variable s:

$$Si = (s,s,\ldots,s) \qquad (1 <= s <= D-1).$$

This general expression follows from the combinatorial nature of the problem at hand. Note that it is a recursive formula: the next higher order of F is defined by a summation of all previous. For example, for the second order problem mentioned earlier, we have for $c1<=c2<=D$:

$$Z = F2(c1,c2,D) = \sum_{s=1}^{D-1} F1(s,s) - \sum_{s=1}^{D-c1} F1(s,s) + c2$$

$$= \sum_{s=1}^{D-1} s - \sum_{s=1}^{D-c1} s + c2 = .5*[-(c1**2)+((2*D+1)*c1)-(2*D)]+c2$$

where the last expression follows using the formula:

$$Q1(t) = \sum_{s=1}^{t} s = .5*(t**2+t)$$

We now turn to some experimental results in gene initiation site classification.

5.0 Experiments in Classification of Translational Sites

5.1 Second Order Separation

To separate gene and nongene initiation sites, each of the four bases A, C, G, and T were given binary values 1000, 0100, 0010 and 0001 respectively. Since each base is uniquely represented in four characters, the dimension of the linear pattern is always four times the number of nucleotides. For these experiments, a positive scaler product between W and a training set signal implied a gene initiation site, a negative value, a nongene site.

The first experiment used the 2nd order algorithm to separate 51 nucleotide strands of mRNA (-30,+20) that were previously found not separable by Stormo's linear model. A total of 124 gene initiation sites were combined with 945 nongene sites to form the training set. Each of these signals was mapped to a 204 element binary pattern (4 bases per nucleotide times 51 nucleotides), which in turn was mapped to the 20911 element 2nd order space.

Training terminated with the desired separation after 22 passes through the 1069 signal training set. The work consumed approximately 7 minutes of VAX 11/780 CPU time running under the VMS operating system. With this, a separation had been realized that had not been feasible using the linear model.

Before going on to a more in-depth analysis of the W that resulted from training, the 2nd order algorithm was "pushed" to train on a shorter 41 nucleotide region (-25,+15) of the same training set signals. Again training was terminated with the desired separation, but this time, 24 passes were required taking 7.5 minutes of VAX 11/780 time. The increased work may be attributed to the lesser degree of freedom given by the shorter signals.

When examined, the W matrix produced from region (-25,+15) training showed heavy gene site correlation with the expected ATG codon in (+00,+02). To graphically display the extent of this correlation, a table of variances is provided (see Tables 5.1-1a, 5.1-1b and 5.1-1c).

The statistical variance is a measure of deviation from mean for a given sample. In this case, the variance is calculated using values from W (-25,+15) corresponding to each base (A, C, G and T) over each of the 41 nucleotide positions. Where the numerical differences between bases are great, the variance is a high positive value indicating that the presence of a base (or set of bases) at the given position is important in the classification.

In Tables 5.1-1, the rows and columns correspond to combinations of 41 nucleotide positions taken two at a time (e.g. 2,2; 2,3; etc.). Looking at Table 5.1-1b, we see a triangle of high variances formed by rows 00-02 and columns 00-02 corresponding to the heavy weight given the ATG codon in the classification. (An examination of W (-25,+15) confirms that the ATG codon is important in classifying gene sites; see Appendix A.) Other regions of high variance are marked and as expected, indicate the importance of the Shine and Dalgarno sequences in the region (-15,-07). Note that the variances do not tell us toward which set the region is important, but only that the region is important in the classification.

Before drawing further conclusions on the results of 2nd order separation, the results of a final experiment in 3rd order separation are presented.

Table 5.1-1a Variance for 2nd Order W (-25,+15)

c2	-25	-24	-23	-22	-21	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11
c1															
-25	4	14	21	15	16	6	19	11	6	15	19	13	15	11	8
-24		5	18	7	15	16	20	8	20	14	13	17	21	12	29
-23			1	11	22	25	10	37*	26	17	28	10	18	17	17
-22				4	20	8	7	12	30*	25	11	12	23	25	11
-21					10	8	20	15	21	17	27	21	13	11	10
-20						2	20	9	23	20	18	5	15	7	23
-19							11	20	5	7	2	15	11	20	16
-18								0	12	24	11	23	24	8	13
-17									7	21	23	9	16	11	11
-16										12	22	11	22	23	16
-15											1	14	9	25	12
-14												8	5	24	22
-13													14	17	27
-12														6	26
-11															7
-10															
-09															
-08															
-07															
-06															
-05															
-04															
-03															
-02															
-01															
00															
01															
02															
03															
04															
05															
06															
07															
08															
09															
10															
11															
12															
13															
14															
15															

Table 5.1-1b Variance for 2nd Order W (-25,+15) Continued

c2	-10	-09	-08	-07	-06	-05	-04	-03	-02	-01	00	01	02
c1													
-25	14	13	16	11	8	21	20	10	24	11	7	6	9
-24	16	21	6	24	20	7	17	17	10	13	5	4	4
-23	8	17	19	14	12	22	14	11	6	16	5	4	2
-22	11	30*	9	17	22	15	9	24	10	10	5	6	3
-21	20	17	18	13	6	14	8	20	17	10	9	10	7
-20	18	10	12	8	26	21	6	9	13	14	13	9	3
-19	14	14	24	28	17	15	10	17	23	11	10	9	8
-18	17	15	33*	15	13	18	18	14	16	15	5	7	5
-17	18	17	16	15	15	28	14	13	25	13	5	6	8
-16	19	11	11	6	15	20	13	12	10	19	7	7	11
-15	11	20	35*	25	5	12	11	13	16	28	7	6	4
-14	7	14	7	10	9	19	22	17	16	8	11	16	6
-13	32*	15	30*	12	10	12	44*	19	10	18	16	9	12
-12	10	15	25	27	38*	9	9	13	10	21	9	10	7
-11	33*	17	16	8	8	14	16	7	20	9	11	22	11
-10	23	39*	11	10	12	23	29	19	15	18	19	31*	14
-09		34*	28	18	20	19	27	14	7	16	19	30*	26
-08			22	22	11	12	10	14	12	15	22	27	13
-07				11	38*	23	5	15	13	10	15	9	7
-06					6	6	15	21	32*	10	7	12	7
-05						1	8	20	19	12	7	5	5
-04							19	17	6	23	26	22	21
-03								16	16	19	18	17	14
-02									5	9	8	10	4
-01										19	18	21	10
00											39*	54*	33*
01												64*	49*
02													19
03													
04													
05													
06													
07													
08													
09													
10													
11													
12													
13													
14													
15													

Table 5.1-1c Variance for 2nd Order W (-25,+15) Continued

c2	03	04	05	06	07	08	09	10	11	12	13	14	15
c1													
-25	21	14	13	12	16	14	28	18	9	25	14	9	13
-24	13	24	6	6	6	9	10	22	8	6	18	22	27
-23	28	7	10	8	18	9	23	22	15	7	15	8	6
-22	16	20	11	10	5	16	14	19	32*	21	28	9	29
-21	23	32*	5	18	19	15	17	10	26	16	34*	24	16
-20	10	18	10	8	10	8	14	10	25	17	12	5	12
-19	14	25	20	6	16	9	8	15	17	27	5	9	14
-18	10	26	18	22	8	15	8	15	10	9	15	12	39*
-17	2	16	15	16	32*	29	12	13	16	8	17	9	12
-16	10	38*	17	13	19	28	12	10	21	8	15	22	12
-15	19	12	20	6	7	9	8	7	9	11	7	24	18
-14	13	24	8	19	10	13	29	7	12	20	7	24	30*
-13	31*	16	20	9	13	17	17	19	10	9	16	9	35*
-12	9	9	7	20	17	12	16	18	8	18	20	14	5
-11	29	23	11	6	7	10	8	6	9	22	8	12	17
-10	8	18	18	17	31*	16	20	8	8	19	9	28	13
-09	17	17	23	21	24	18	8	8	14	17	10	14	6
-08	9	10	35*	20	13	16	27	19	20	10	18	17	28
-07	15	30*	7	15	8	22	19	12	16	24	12	26	20
-06	14	12	15	29	15	18	22	9	17	6	27	12	30*
-05	10	17	8	10	9	9	11	26	22	8	12	16	12
-04	7	31*	10	21	8	21	8	11	5	10	13	21	13
-03	20	25	31*	14	14	20	17	19	22	11	9	18	12
-02	6	18	19	24	10	18	14	22	19	5	17	26	12
-01	13	15	22	14	17	19	16	15	15	20	13	18	15
00	6	16	11	7	7	6	5	9	11	15	13	9	10
01	7	10	15	15	6	6	17	12	16	18	11	9	9
02	6	9	8	9	2	2	3	7	11	4	4	12	6
03	2	22	24	21	22	9	21	27	21	14	9	15	5
04		20	36*	22	29	15	26	33*	17	10	10	9	12
05			11	13	9	8	15	28	7	10	18	12	14
06				14	48*	21	8	13	16	10	29	9	19
07					6	30*	21	18	16	9	23	31*	20
08						2	15	15	12	6	13	12	15
09							4	28	3	9	43*	24	19
10								10	15	16	24	13	7
11									15	27	18	17	10
12										7	21	20	12
13											9	13	21
14												14	12
15													8

5.2 Third Order Separation

As discussed in Section 2, the codon is the basic unit within mRNA strands from which protein is built by ribosomes. This being the case, it makes sense to extend work done in nonlinear methods to consider combinations of nucleotides taken three at a time.

For this experiment, the region (-15,+05) was considered using the same training set as in the 2nd order problem. Each of the 21 nucleotide signals were represented as 102,341 element points in the 3rd order space. Convergence upon the desired functional W was achieved in 13 passes though the training set (1069 signals) taking about 10 minutes of VAX 11/780 CPU time. The greater amount of time required for each pass is attributed to the greater complexity of the 3rd order transformation derived from the equation given in Section 4.0.

The 3rd order variance table is considerably larger than Table 5.1-1. Only a relevant portion of this table is presented here (see Table 5.2-1). The table shows all combinations arising from the -11 strand position. The rows and columns correspond to the remaining indices defining valid combinations (e.g. (-11,-10,-08)).

Immediately apparent is the equality of values along the top row and the diagonal. These variances are for the second order terms and are equal because the existence of a diagonal combination (e.g. (-11, -10,-10)) necessarily implies the existence of the corresponding top row combination (e.g. (-11,-11,-10)). During training, each W location is weighted exactly the same, and hence the resulting W values (and their variances) are identical.

Also apparent is the generally higher values for variances associated with the second order terms. This is because these terms occur more frequently in the training set and thus appear more frequently during error correction. This leads to values of generally greater absolute value which in turn produces higher variances. To normalize this affect, the mean of variance over the first, second and third order terms was computed:

MV1 = 4.5
MV2 = 9.0
MV3 = 4.0

Many of the variances in Table 5.2-1 are well above these means. In fact, this tableau was selected because for each order of terms, the highest variance was also the highest variance over the remaining 20 tableaus. These values are so indicated with a trailing asterisk.

Table 5.2-1 Variance for 3rd Order W (-15,+05)

c1					-11						
c3	-15	-14	-13	-12	-11	-10	-09	-08	-07	-06	-05
c2											
-11					9*	26*	14	6	3	7	10
-10						26*	6	4	4	5	7
-09							14	6	4	5	5
-08								6	4	4	4
-07									3	4	6
-06										7	4
-05											10
-04											
-03											
-02											
-01											
00											
01											
02											
03											
04											
05											

c1										
c3	-04	-03	-02	-01	00	01	02	03	04	05
c2										
-11	11	4	11	8	7	16	5	11	6	3
-10	8	4	5	6	7	10*	7	4	5	5
-09	6	4	4	5	4	5	5	5	4	4
-08	5	3	5	4	3	4	2	4	5	4
-07	2	3	3	4	1	2	1	4	4	3
-06	4	4	4	5	3	3	2	5	4	4
-05	5	5	5	6	3	3	3	5	4	4
-04	11	2	4	5	3	4	5	5	6	3
-03		4	5	4	2	2	2	4	6	4
-02			11	4	3	4	3	4	5	4
-01				8	3	3	3	4	3	4
00					7	5	2	6	2	1
01						16	5	4	3	3
02							5	3	2	1
03								11	6	5
04									6	4
05										3

From this analysis the conclusion may be drawn that the -11 tableau in general, and the combinations (-11,-11,-11), (-11,-11,-10) and (-11,-10,+01) in particular, are somehow very important in the classification of the training sites. Also found to be important (with similar high variances), are the tableaus associated with positions -12, +00 and +01 (see Appendix B).

5.3 Conclusions

That the 3rd order -11 and -12 tableaus proved to be important is not surprising in that this region is the center of the highly correlated Shine and Dalgarno sequence. What is surprising is that the (+00,+02) region, so important in 2nd order classification, was given a relatively low 3rd order score. An explanation for the low variance is that many nongene sites in the training set had the ATG codon in the (+00,+02) region. The question still remains as to why the 2nd order variances were not similarly low.

One explanation might be that the added degree of freedom in 3rd order separation allowed for the desired functional to be realized taking into account only the Shine and Dalgarno region (which also was important in the 2nd order case; see Table 5.1-1), ignoring the "less clear" (+00,+02) region with the conflicting ATG presence. The more constrained 2nd order separator made more passes through the training set (24 against 13 in the 3rd order case), and perhaps in that time had to give some added importance to the (+00,+02) region in order to converge upon W. If this is the case, constraining the 3rd order separator by increasing the size of the training set may well serve to settle the question.

It is interesting that the highest third order variance is associated with the (-11,-10,-01) combination which happens to define an interaction between the Shine and Dalgarno region and the central ATG codon. Why this combination alone proved important (other interactions with the central region were notably low), is again probably a matter of degrees of freedom. Nevertheless, that the combination was given the highest variance over the 3rd order table suggests that the presence of the central ATG codon by itself may not be as important as its interaction with other specific regions on the strand. In further support of this conjecture, both Tables 5.1-1 and 5.2-1 show very steep variance gradients over the (+00,+02) region suggesting that the central region is important only when it interacts with other specific regions (such as the Shine and Dalgarno area).

6.0 Acknowledgments

I would like to thank Gary Stormo for his time and patience in familiarizing me with the area of molecular biology which is the subject of this paper. Also, his pioneering in the use of a Perceptron to recognize gene initiation sites was the foundation for this work in nonlinear methods. Jeff Haemer was of enormous assistance with regard to the use of the language DELILA used to extract from a data base the gene and nongene sites used in training. Finally, I would like to thank A. Ehrenfeucht of the University of Colorado for providing the inspiration to pursue this area of artificial intelligence.

References

1. H. Block, B. Knight and F. Rosenblatt, "Analysis of a Four Layer Series-Coupled Perceptron," Reviews of Modern Physics 34, 135 (1962).

2. K. Schlesinger and P. Groves, "Psychology A Dynamic Science," 39-44 (W.C. Brown Company, 1976).

3. G. Stormo, "Computer Aided Characterization of Translational Initiation Sites - E.Coli," Doctoral Thesis, University of Colorado (1981). (A new paper is forthcoming.)

4. H. Greenberg and A. Konheim, "Linear and Nonlinear Methods in Pattern Classification," IBM Journal of Research and Development 8, 299 (July, 1964).

Appendix A

The following is a portion of the 2nd order W realized over region (-25,+15). Note the high positive values (for gene initiation site correlation) associated with the ATG codon in position (+00,+02).

The bases in the first row of each tableau are the first of a pair. The remaining bases in the combination are in the second row, four for each first row base. The weights given the combinations are listed beneath the second row base for each relevant nucleotide position.

00	A				C				G				T			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
00	6*					-8					-2					-10
01	-9	-6	-4	25*	0	0	0	-8	-1	0	-1	0	-1	-1	-1	-7
02	-3	-4	19*	-6	0	0	-8	0	-1	-1	0	0	-1	-1	-7	-1
03	4	-4	3	2	-5	0	-3	0	1	0	-1	-2	-2	-1	-4	-3
04	6	10	-4	-7	-4	-3	0	-1	0	-2	1	-1	-2	-5	-1	-2
05	2	1	-6	8	-3	-2	0	-3	3	0	-3	-2	-3	-3	0	-4
06	7	2	0	-4	-3	0	-3	-2	-3	1	1	-1	-3	-2	-3	-2
07	3	3	4	-5	-4	-2	-2	0	2	0	-1	-3	-3	-4	-3	0
08	-2	0	6	1	0	-1	-3	-4	-1	-1	-2	2	-2	-3	-4	-1
09	-1	-1	3	4	-3	-1	-2	-2	-1	-1	-1	1	-2	0	-3	-5
10	7	2	-5	1	-3	-1	0	-4	-3	-2	0	3	0	-3	-3	-4
11	4	1	-7	7	-5	0	-1	-2	1	-1	-2	0	-4	-1	0	-5
12	10	3	-8	0	-6	-2	0	0	-1	-2	1	0	-5	0	-1	-4
13	8	-4	-6	7	-3	0	-1	-4	-2	0	1	-1	-4	-1	-2	-3
14	2	2	-6	7	-4	-1	-1	-2	0	-3	0	1	-1	-4	-1	-4
15	5	0	5	-5	-7	0	-1	0	0	-4	-1	3	-1	-4	-3	-2

01	A				C				G				T			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
01	-11					-7					-6					10*
02	-1	0	-10	0	-1	0	-6	0	0	-1	-4	-1	-3	-5	24*	-6
03	-5	-2	-2	-2	-2	-1	-2	-2	-2	-1	-1	-2	7	-1	0	3
04	-2	-6	-1	-2	-4	0	-1	-2	-1	0	-2	-3	7	6	0	-4
05	-5	0	-3	-3	-5	-1	-1	0	-2	0	-1	-3	11	-3	-4	5
06	-2	-3	-5	-1	-3	-3	-1	0	-4	0	-2	0	7	7	3	-8
07	-1	-2	-2	-6	-3	-3	-1	0	-1	0	-3	-2	3	2	4	0
08	-5	-1	-3	-2	-3	-3	-1	0	-2	-4	0	0	5	3	1	0
09	0	-1	-1	-9	-3	-2	-1	-1	0	0	-2	-4	-4	0	1	12
10	-1	-3	-2	-5	0	-1	-1	-5	-1	0	-1	-4	3	0	-4	10
11	-2	-1	-2	-6	0	-4	-1	-2	-2	-1	0	-3	0	5	-7	11
12	-3	-6	0	-2	-4	-1	-2	0	-6	0	0	0	11	6	-6	-2
13	-4	-2	-1	-4	0	-2	-1	-4	-2	0	-2	-2	5	-1	-4	9
14	-3	-4	-2	-2	0	-3	-1	-3	-1	-1	-2	-2	1	2	-3	9
15	-4	-3	-2	-2	-4	0	-3	0	-2	-2	0	-2	7	-3	5	0

02	A					C					G					T
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
02	-5					-6					4*					-7
03	-2	-1	0	-2	-3	-2	0	-1	5	-1	-5	4	-2	-1	0	-4
04	-1	-2	-2	0	-1	-4	-1	0	3	7	0	-7	-1	-1	-1	-4
05	-1	-1	-1	-2	-2	-1	-1	-2	5	-1	-7	6	-3	-1	0	-3
06	-1	-2	-1	-1	-2	-1	-3	0	5	6	-1	-7	-4	-2	0	-1
07	-1	-1	-1	-2	-2	-1	-1	-2	3	1	1	-2	-2	-2	-1	-2
08	-2	0	-2	-1	-2	-1	-1	-2	0	-1	1	3	-1	-3	-1	-2
09	-4	0	0	-1	-2	0	-1	-3	1	-2	1	3	-2	-1	-3	-1
10	0	-1	-2	-2	-4	-1	0	-1	6	1	-6	2	-1	-3	0	-3
11	-1	-1	-1	-2	-1	-2	0	-3	2	3	-9	7	-4	-1	0	-2
12	-1	-2	-1	-1	-4	0	-1	-1	5	2	-3	-1	-2	-1	-3	-1
13	-2	-1	-1	-1	-1	-1	-2	-2	5	0	-4	2	-3	-3	-1	0
14	0	-1	0	-4	-2	-2	-1	-1	0	-3	-5	11	-1	0	-2	-4
15	-2	0	-2	-1	-3	-1	-1	-1	3	-6	5	1	-1	-1	-2	-3

Appendix B

The following is an additional portion of the 3rd order W variance table. Included are tableaus which exhibit importance with regard to classification by virtue of the high variance values (with asterisks).

				-12							
	-15	-14	-13	-12	-11	-10	-09	-08	-07	-06	-05
-12				2	25*	12	11	14	10	24	10
-11					25*	9*	6	7	7	7	6
-10						12	4	4	4	5	5
-09							11	5	4	7	4
-08								14	6	8	5
-07									10	6	5
-06										24	5
-05											10
-04											
-03											
-02											
-01											
00											
01											
02											
03											
04											
05											

	-04	-03	-02	-01	00	01	02	03	04	05
-12	4	13	12	9	2	3	3	9	7	6
-11	4	6	6	5	6	7	6	6	5	4
-10	5	5	3	4	4	4	4	4	7	5
-09	5	5	5	3	3	4	4	4	5	7
-08	5	7	4	5	4	5	5	4	5	4
-07	3	5	6	4	3	3	2	4	4	3
-06	5	5	7	5	5	7	7	7	7	5
-05	3	4	5	4	3	3	3	4	6	4
-04	4	3	5	4	2	2	2	4	4	3
-03		13	6	5	4	4	4	6	5	5
-02			12	5	4	3	3	3	5	4
-01				9	2	3	3	3	4	4
00					2	1	1	3	3	3
01						3	2	2	2	3
02							3	3	2	2
03								9	4	5
04									7	5
05										6

```
   00
   00 01 02 03 04 05
00  9* 7* 3  5  6  6
01     7* 5* 2  2  2
02        3  1  2  2
03           5  4  4
04              6  3
05                 6
```

```
   01
   01 02 03 04 05
01  4  9* 4  4  4
02     9* 2  2  2
03        4  4  3
04           4  3
05              4
```

NONLINEAR CLASSIFICATION
OF
TRANSLATIONAL INITIATION
SITES

by

William W. Brown
Department of Computer Science
University of Colorado
Boulder, Colo. 80309

CU-CS-211-81 October 1981

Completed August 1981 as CS-950 Graduate Independent Study for the M.S. degree in computer science.

NONLINEAR CLASSIFICATION
OF
TRANSLATIONAL INITIATION
SITES

by

William W. Brown
Department of Computer Science
University of Colorado
Boulder, Colo. 80309

CU-CS-211-81 October 1981

Completed August 1981 as CS-950 Graduate Independent Study for the M.S. degree in computer science.