22. Tremblay, J.-P. and Sorenson, P. G., *An Introduction to Data Structures with Applications*, McGraw-Hill, New York, 1984.
23. Ornstein, R. L. and Fresco, J. R., Correlation of $T_m$, sequence, and $\Delta H$ of complementary RNA helices and comparison with DNA helices, *Biopolymers*, 22, 2001, 1983.
24. Waterman, M. S. and Smith, T. F., Rapid dynamic programming algorithms for RNA secondary structure, *Adv. Appl. Math.*, 7(4), 455, 1986.
25. Zuker, M. and Stiegler, P., Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, *Nucl. Acids Res.*, 9, 133, 1981.
26. Mills, D. R., Kramer, F. R., and Spiegelman, S., Complete nucleotide sequence of a replicating RNA molecule, *Science*, 180, 916, 1973.
27. Stiegler, P., Carbon, P., Zuker, M., Ebel, J.-P., and Ehresmann, C., Structural organization of the 16S ribosomal RNA from *E. coli*. Topography and secondary structure, *Nucl. Acids Res.*, 9, 2153, 1981.
28. Vary, C. P. H. and Vournakis, J. N., RNA structure analysis using methidiumpropyl-EDTA.Fe(II): a base-pair specific RNA structure probe, *Proc. Natl. Acad. Sci. U.S.A.*, 81(22), 6978, 1984.
29. Riesner, D., Colpan, M., Goodman, T. C., Nagel, L., Schumacher, J., Steger, G., and Hofmann, H., Dynamics and interactions of viroids, *J. Biomol. Struct. Dyn.*, 1, 669, 1983.
30. Zuker, M., RNA folding prediction: the continued need for interaction between biologists and mathematicians, *Lect. Math. Life Sci.*, 17, 1986.
31. Waterman, M. S., Sequence alignments in the neighborhood of the optimum with general application to dynamic programming, *Proc. Natl. Acad. Sci. U.S.A.*, 80, 3123, 1983.
32. Waterman, M. S. and Byers, T. H., A dynamic programming algorithm to find all solutions in a neighborhood of the optimum, *Math. Biosci.*, 77, 179, 1985.
33. Williams, A. L., Jr. and Tinoco, I., Jr., A dynamic programming algorithm for finding alternative RNA secondary structures, *Nucl. Acids Res.*, 14(1), 299, 1986.
34. Stiegler, P., Carbon, P., Ebel, J.-P., and Ehresmann, C., A general secondary structure model for procaryotic and eucaryotic RNAs of the small ribosomal subunits, *Eur. J. Biochem.*, 120, 487, 1981.
35. Sankoff, D., Simultaneous solution of the RNA folding, alignment and protosequence problems, *SIAM J. Appl. Math.*, 45, 810, 1985.
36. Turner, D. H., personal communication.



Chapter 8

# CONSENSUS METHODS FOR FOLDING SINGLE-STRANDED NUCLEIC ACIDS

Michael S. Waterman

## TABLE OF CONTENTS

## I. INTRODUCTION

As the preceding chapter[1] has explained, the structure of single-stranded RNA macromolecules is crucial to the functioning of an organism. While it has recently become routine to directly read the primary structure of these molecules by sequencing techniques, the deduction of secondary and tertiary structure is much less straightforward. The secondary structure of DNA is well known: DNA is double-stranded according to the familiar Watson-Crick rules. Double-stranded DNA has alternate double helical structures. The classic B- and A-forms are both right-handed helixes while the Z-form is left-handed.[2,3] RNA has base-pairing rules corresponding to those for DNA; T in DNA is replaced by U so that base A pairs U (A*U) and base G pairs C (G*C). The pair G*U is usually added to this list. The fact that RNA occurs frequently as single-stranded often makes the secondary structure of RNA difficult to determine. Segments of the sequence will form base pairs between them, and the prediction of the resulting structure is a difficult task. Obviously the resulting structure — the folded molecule — is highly dependent on the specific linear sequence of the RNA.

It is quite surprising, to a mathematician at least, that biologists have been so successful at predicting some important secondary structures. In fact, the first primary (linear) sequence of a tRNA (transfer RNA) appeared in 1965[4] along with the cloverleaf form of secondary structure. This turned out to be the correct structure and has been verified by X-ray crystallography.[5] Other than by guessing or inspection, there seem to be two major techniques for prediction of secondary structure: the minimum energy method and the comparative method. The previous chapter gives an extensive treatment of the important minimum energy method, which utilizes dynamic programming. After briefly discussing the minimum energy approach we turn to the main topic of this chapter, comparative or consensus analysis of folding.

In an important paper Tinoco et al.[6] proposed assigning free energies to the components of secondary structure — the various base pairs, end loops, bulges, interior loops, and multibranch loops — and then finding the minimum free energy secondary structure. To accomplish this task they presented the base pairing matrix for an RNA, which is the analog of the dot matrix for sequence matching. One difficulty with fully implementing their proposal is the huge number of possible secondary structures. The number of configurations has been studied,[7] and it was found that for sequences of length 150, allowing end loops of two bases or more, there are $1.22 \times 10^{54}$ possible secondary structures. Now this number counts all conceivable structures, and the base pairing of a given sequence reduces the number somewhat, but the point remains. There are too many candidate structures to simply consider them all and take the one with minimum free energy.

In 1978, two dynamic programming methods were proposed to solve this problem. Waterman[8] and Waterman and Smith[9] used general energy functions and, in an iterative fashion, built up the complexity of the optimal structures. Nussinov et al.[10] maximized the number of base pairs in a single pass algorithm. The advantages of both of these methods have been combined into a useful, efficient algorithm described in the preceding chapter.[1]

Some of the shortcomings of the minimum energy methods are (1) the lack of precise knowledge about the energy functions themselves, (2) the large amount of computer time required, and (3) the inability of the current algorithms to handle many sequences simultaneously. Item 3 is really a subcategory of 2, since computer time and storage is the main difficulty in 3. Sankoff[11] has an algorithm to simultaneously fold and align several sequences; three sequences seem to be an upper limit, but the ability to both fold and align several sequences is an important problem. To overcome these difficulties, it is instructive to take a careful look at some of the successful work of biologists who study these problems.

In a remarkable 1969 paper, Levitt[12] obtained a cloverleaf model that fit all the 14 tRNA sequences known at that time. He almost certainly obtained his consensus structure by



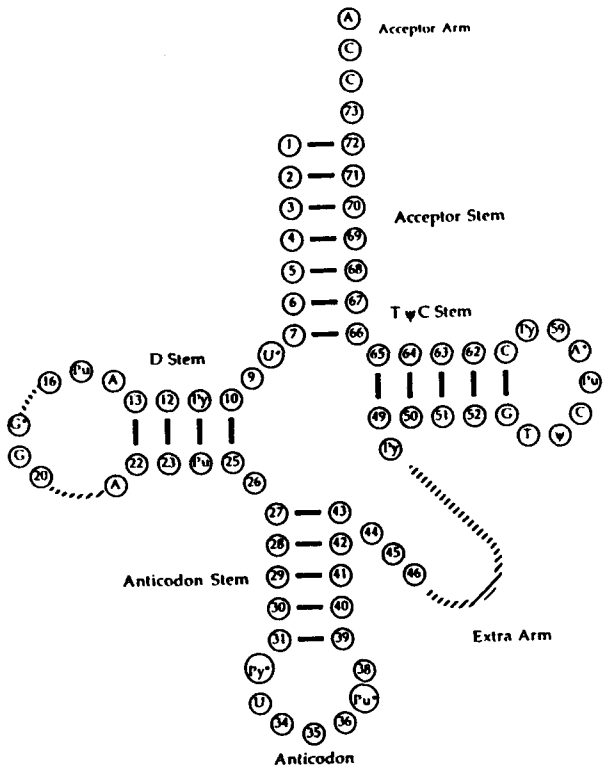



FIGURE 1. The cloverleaf structure of tRNA. The nucleotide positions are numbered, or whenever a position is conserved, the identity is indicated. Dashed lines indicate nucleotides that may or may not be present in any given molecule.

arranging the sequences by hand into an alignment in which helixes and homologies (identical bases) were represented. Later, the structure was largely confirmed by crystallography.[5] Essentially the same pattern of helix and homology is shown in Figure 1 where a so-called universal cloverleaf is shown.[13] The over 300 known tRNA sequences[14] fit this general structure, and we now discuss it in some detail.

The meaning of Figure 1 is that all known tRNA sequences can be arranged into an alignment with positions labeled as in the figure. The appearance of a base pair, between 1 and 72, for example, means that whatever the identity of the bases labeled 1 and 72, they form a base pair. The sequences end in CCA; this triplet has been conserved, independent of any base pairing. Actually the "universals" are in some positions violated by 5 to 10% of tRNA sequences. The variation in sequence length makes structure prediction and/or alignment a fascinating and difficult problem. The D arm varies in length by up to 4 bases, while the so-called extra arm varies in length by up to 18 bases. Therefore, the alignment

is not simply given by shifts of the primary sequence, but must be found by inserting gaps into the sequences. Notice the large number of conserved bases in the interior of the sequence. Positions 53, 54, 55, and 56 are GUUC, the longest conserved sequence in the structure. In total, 14 of the bases of tRNA are conserved in the alphabet {A,U,G,C}, while 8 more are conserved in the Purine = Pu and Prymadine = Py alphabet, {Pu, Py}.

Let us take a brief look at the magnitude of the problem of sequence alignment, which we must solve to put the sequences into correspondence. The tRNA sequences can differ in length by as much as 20 bases. If there are R sequences and we want to look at them in all possible arrangements, not allowing gaps within the sequences, there are a minimum of $(20 + 1)^R$ possible arrangements. If R = 14, as was Levitt's situation, $(21)^{14} \approx 3.24 \times 10^{18}$. If R = 32, as in the example of this chapter, $(21)^{32} \approx 2.05 \times 10^{42}$. If R = 150, a reasonable number of tRNA sequences, $(21)^{150} \approx 2.15 \times 10^{198}$. In none of these cases is it possible to exhaustively consider all alignments on any modern computer. Allowing gaps from 1 to 20 letters only increases these numbers by many orders of magnitude. Even with no gaps and only R = 14 sequences, a direct approach to alignment is computationally hopeless. Clearly, the approach of considering all alignments individually is not feasible, even for the smallest cases of interest.

The approach of inferring structure by common (conserved) features, helixes, or homologous bases, has come to be known as the comparative or phylogenetic method. Features in rRNA (ribosomal RNA) essential to organisms must be conserved in evolution, and it is hoped that these features will be recognized as common in the sequences studied. For example, the CCA at the 5′ end of tRNA is involved in the interaction between tRNA and the amino acids. By locating CCA in a tRNA sequence, we obtain valuable information about the location of the acceptor stem. Levitt's approach was based on these ideas.

After tRNA, the next RNA molecule for which this method was used was 5S rRNA, which is about 120 bases in length. Fox and Woese[15] and Nishikawa and Takemura[16] solved the structure after many attempts by investigators using other approaches. There is no crystallographic data for this molecule, but biochemical and physiochemical evidence support the structure. (See Waterman[17] for a consensus approach to folding 34 5S rRNA sequences on a computer.) The difficulties of unequal sequence lengths exist with 5S sequences also. A study of Trifonov and Bolshoi[18] studies 5S folding by a related method which has some drawbacks. First, the sequences must be aligned, a difficult problem in itself. Then the base pair matrix[9] for each sequence is obtained. The matrixes are summed and possible helixes appear as dark, antidiagonal regions. The methods presented in this chapter directly consider the helixes.

The next larger rRNA, 16S, and 16S-like molecules posed new difficulties for investigators, due in part to the greater sequence length of approximately 1540 bases. As in 5S sequences, bulge loops, interior loops, and noncanonical base pairs (those other than A*U or G*C) appear in 16S structures. As with 5S structure, 16S structure has been solved by the comparative method. The work was mainly done by three groups, and it is this work, notably that of Woese and Noller,[19,20] that provided the motivation and inspiration for this chapter (see also References 21, 22, and 23). The model of Woese and Noller and collaborators[24] for 16S rRNA of *Escherichia coli* is shown in Figure 2: These authors describe their approach[19] as progressing with alignment of 16S-like rRNA sequences in parallel with the development and testing of the secondary structure model. Preliminary alignments are then used to identify obvious primary and secondary structure features; these patterns are used as the basis for refinement of the sequence alignment. The procedure is then iterated, and new sequence information incorporated as it is obtained.

It is the goal of this chapter (1) to make some of the Noller-Woese procedures explicitly defined so that other groups can see exactly what the corresponding computer searches are, (2) to find efficient computer methods to perform the searches, and (3) to give some estimates



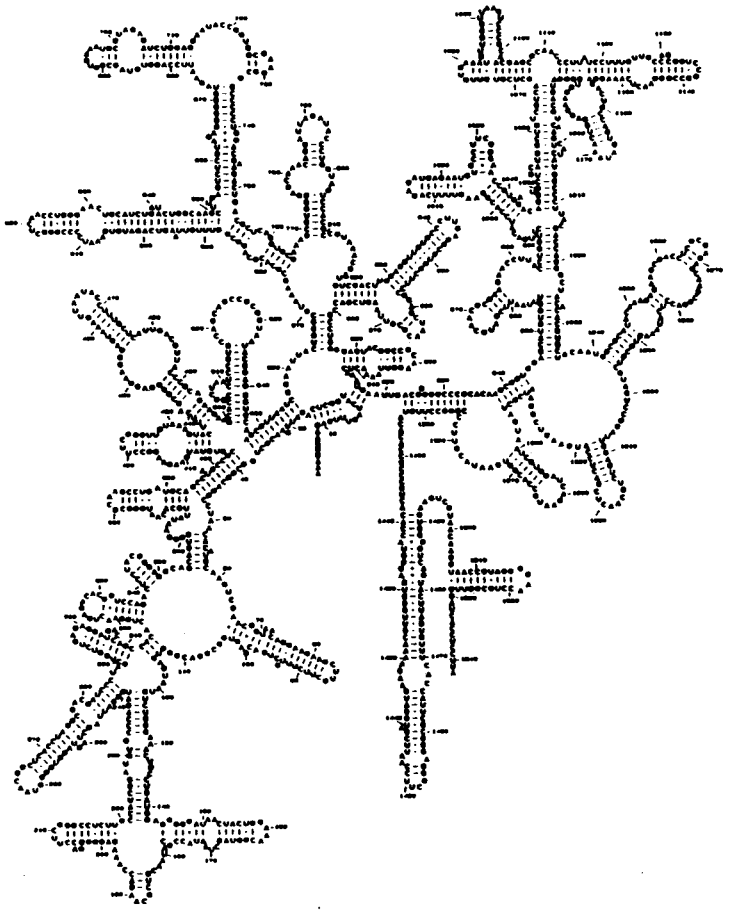


FIGURE 2. Secondary structure of *E. coli* 16S RNA. (Taken from Moazed, D., Stern, S., and Noller, H., *J. Mol. Biol.*, 187, 399, 1986. With permission.)

of statistical significance. The philosophy beneath the comparative method is that important features, such as specific bases or helixes, have been conserved over the course of RNA evolution and that these conserved features are still utilized by the organism. Our task is to make this approach into an algorithm.

For illustrative purposes we will study the set of 32 *E. coli* tRNA sequences whose names appear in Table 1. Since some of the sequences have the extra arm, the variation in length is from 74 to 93 bases. Several other interesting features arise. For example, the acceptor stem is shifted one base from the "universal" structure in the sequence of histidine tRNA. While tRNAs are the shortest RNA sequences we have discussed, they will serve well for

**Table 1**
**THE 32 *E. COLI* tRNA SEQUENCE NAMES WITH THE GenBank ABBREVIATIONS**

| | |
|---|---|
| ala tRNA; | ECOTRA1A |
| ala tRNA; | ECOTRA1B |
| cys tRNA; | ECOTRC |
| asp tRNA; | ECOTRD1 |
| glu tRNA; | ECOTRE1 |
| glu tRNA; | ECOTRE2 |
| phe tRNA; | ECOTRF |
| gly tRNA; | ECOTRG1 |
| gly tRNA; | ECOTRG2 |
| gly tRNA; | ECOTRG3 |
| his tRNA; | ECOTRH1 |
| ile tRNA; | ECOTRI1 |
| ile tRNA; | ECOTRI2 |
| lys tRNA; | ECOTRK |
| leu tRNA; | ECOTRL1 |
| leu tRNA; | ECOTRL2 |
| leu tRNA; | ECTORL5 |
| initiator met tRNA; | ECOTRMF |
| met tRNA; | ECOTRM |
| asn tRNA; | ECOTRN |
| gln tRNA; | ECOTRQ1 |
| gln tRNA; | ECOTRQ2 |
| arg tRNA; | ECOTRR1 |
| arg tRNA; | ECOTRR2 |
| ser tRNA; | ECOTRS1 |
| ser tRNA; | ECOTRS3 |
| thr tRNA; | ECOTRTACU |
| val tRNA; | ECOTRV1 |
| val tRNA; | ECOTRV2A |
| val tRNA; | ECOTRV2B |
| trp tRNA; | ECOTRW |

illustration. Both 16S and 23S are too long with which to easily illustrate the algorithms, although even with their greater length they are still computationally feasible. 16S has about 1540 bases while 23S[25] has about 2500 bases. As will be seen, tRNAs are quite suitable for our purposes, being of manageable size and of sufficient difficulty of folding. Dynamic programming methods are reputed to fold approximately half of the tRNA sequences into a cloverleaf structure. Consensus methods, as will be seen, fold all *E. coli* tRNAs into the correct cloverleaf shape. In addition, tertiary structure can be studied by the same methods.

## II. ALIGNMENT BY MATCHES

In earlier chapters on alignment by Karlin[26] and Waterman,[27] the authors discussed finding statistically significant matches between sequences. The motivation for that work is the hope that statistically significant matches between sequences will be biologically significant. In the setting of this chapter, there are several conserved (or "invariant" or "universal") sequences in tRNA. In Figure 1, we notice that the longest universal pattern is GTTC. In our data we will see that with a few exceptions this pattern is contained in a longer five letter pattern GGTTC, beginning at base 52 in the figure. The acceptor arm pattern of CCA is also present in all sequences. Several other one and two letter patterns can be seen on examining Figure 1.



There are several reasons for being interested in these invariant patterns. Their presence has been conserved over evolutionary time, and this gives us some reason to believe they are essential to the functioning of tRNA. Our basic reason for study of these data sets is to deduce structure, function, and evolution of the macromolecules. As an aid in deducing the structure, then, finding significant invariant patterns could be essential in deducing the correct alignment of the data set. While no shifting is necessary to find GGTTC if the sequences are already aligned on their right ends, such shifting into the correct alignment could allow us to locate other smaller patterns, as well as base pairing which might otherwise be undetectable. This turns out to be the case with 16S RNA.[20]

It is now time to ask some hard questions. What is the basis for concluding that such patterns are universal? The flaw in looking at a data set until a pattern is seen and then concluding it is significant has often led scientists to incorrect conclusions. Exactly what search is performed to find these patterns? Are other significant patterns missed in Figure 1? How is statistical significance to be estimated? These important questions are addressed in the remainder of this section.

### A. Finding Matches

In the search for matches, our program simply finds all patterns of a specified length that occur at or above a present frequency in a specified section (column "a" to column "b") with the sequences arranged in some alignment. Even if a sequence, GTTC for instance, occurs several times within a single sequence, it is only counted once. For small data sets such as the tRNAs, to find k-letter (k-mer) repeats it is sufficient to make a table of all k-mers occurring in the sequence ($k \leq 9$) with their frequencies and then check to see which occur at the required frequency. The search for k-mer repeats can evidently be performed in time proportional to the number of letters (N say) in all sequences with storage bounded by $O(4^k)$. If the common sequences are longer, then the techniques of hashing allow the search to be done in NlogN time (see Martinez[28] for a useful algorithm to find repeats in molecular sequences by hashing). Thus it is seen that the search for common patterns is not computationally difficult in these problems. In these cases, we are interested in exact matches only.

### B. Statistical Significance

#### *1. The Log(n) Distribution*

Estimates of statistical significance of matches are more difficult than finding the matches and were not well understood until recently. The model we study here is R sequences of length N that have iid (independent, identically distributed) letters. The event for which we calculate significance levels is that of finding a pattern of length k common to L of R sequences. First we present results for the case $N \to \infty$. In Chapter 3, the case of $R = L = 2$ is discussed with extensions to imperfect matchings, while in Chapter 6, extensions to larger R and L are given. We give the simplest of these extensions here. Let $p = P(X_1 = X_2 \ldots = X_k)$ where $X_i$, the letters in the sequences, are iid. If the alphabet has four equally likely letters, $p = 4(1/4)^k = (1/4)^{k-1}$. Let M(R,L) = length of the longest pattern common to at least L of the R sequences. Then,

$$E(M(R, L)) \approx \log(\binom{R}{L}N^{L-1}) + \log(1 - p) + \gamma\log(e) - 1/2$$

and

$$\mathrm{Var}(M(R, L)) = \sigma^2 \approx (\pi \log(e))^2/6 + 1/12$$

where

$$\log = \log_{1/q} \quad \text{and} \quad \gamma \approx .0577$$

For the R = 32 sequences of length ≈ 75 = N, take L = 32 as well. If the letters are equally likely, $p = (1/4)^{31}$ and

$$E(M(32, 32)) \approx 3.1144 \ldots + 0.0134 \ldots - 0.5$$

$$= 2.6278 \ldots$$

and

$$\sigma^2 = 0.0842 \ldots$$

with

$$\sigma \approx 0.2902 \ldots$$

Therefore, 4 = k is almost 5 standard deviations above E(M), and this k is achieved in our sequences by the word GTTC.

*2. The Binomial Distribution and Large Deviations*

A glance at the locations of the pattern GTTC in the data set aligned as shown in Figure 5 brings up an interesting second question. The pattern occurs perfectly aligned in the figure as well as in some other locations. What is the significance of such a pattern of occurrence with little shifting? That is, in the case $N << \infty$. In Figure 5, we could take k = W = 4 and discover the pattern, where W is the width of the window in which the search is being performed.

For ease of exposition, take the case of equally likely letters of RNA. Then, if w is a k-letter sequence,

$$\alpha = P(w \text{ occurs in } N \text{ letters})$$

$$= P\left(\bigcup_{i=1}^{W-k+1} \{w \text{ starts at } i\}\right) \leq (W - k + 1)(1/4)^k$$

The probability that w occurs in exactly L of R sequences is given by the binomial probability

$$\binom{R}{L}\alpha^L(1-\alpha)^{R-L} \approx \binom{R}{L}(W - k + 1)^L(1/4)^{kL}(1 - (W - k + 1)(1/4)^k)^{R-L}$$

and, summing over all $4^k$ possible w, the desired probability is

$$\binom{R}{L}(W - k + 1)^L(1/4)^{kL-1}(1 - (W - k + 1)(1/4)^k)^{R-L}$$

For small R and L, this formula can be directly used to estimate significance. Otherwise we use the large deviations theory described next.

Now define $\beta = L/R$. If $\beta \leq \alpha$ and R is large, the strong law of large numbers assures us that we will have approximately $\alpha R \geq \beta R = L$ occurrences. Otherwise, when $\beta > \alpha$, a bound for the probability of a k-letter word common to at least L of R sequences is given by the large deviations estimate.[29-31]

The estimate is given by



**Table 2**
**ESTIMATES OF STATISTICAL SIGNIFICANCE FOR APPEARANCE OF SOME k-LETTER WORD IN SOME WINDOW POSITION AND IN AT LEAST L OF R = 32 SEQUENCES OF LENGTH N = 75**

| k | L | W | α | β | H(α,β) | p |
|---|---|---|---|---|---|---|
| 4 | 20 | 50 | 0.184 | 0.625 | 0.474 | $1.73 \times 10^{-3}$ |
| 4 | 24 | 75 | 0.281 | 0.75 | 0.472 | $7.14 \times 10^{-5}$ |
| 6 | 10 | 75 | 0.171 | 0.313 | 0.662 | $2.55 \times 10^{-4}$ |
| 7 | 7 | 75 | 0.004 | 0.188 | 0.547 | $4.17 \times 10^{-4}$ |

$$4^k \exp\{-RH(\alpha, \beta)\}$$

where $H(\alpha,\beta) = \beta\log\beta/\alpha + (1 - \beta)\log(1 - \beta)/(1 - \alpha)$. The factor of $4^k$ is to count the number of possible w. This quantity approximates the probability that some k-letter word is common to at least L of R sequences of length N. If a window of width W is placed in all N − W + 1 possible positions, then the estimate becomes

$$(N - W + 1)4^k \exp\{-RH(\alpha, \beta)\},$$

where α = P(w occurs in W letters).

Some sample estimates appear in Table 2. Finding some four letter word common to 24 of 32 sequences of length N = 75 will only happen with the probability of $7.15 \times 10^{-5}$. In our sequences, we find the word w(=GTTC) in all 32 sequences perfectly aligned, a highly unlikely event!

## III. ALIGNMENT BY BASE PAIRING

While alignment by matches common to many sequences is a very useful procedure, the striking feature of rRNA data sets is commonality of base pairing. The most conserved features in Figure 1 are not conserved letters, but conserved base pairs (bp) or helixes. The aminoacyl stem is a helix of 7 bp, while the TψC stem and the anticodon stem have 5 bp, and the D stem has 4 bp.

All of the invariant helixes are, in our data set, composed of differing sequences. The common feature is that a helix of the required length can be formed with a relatively small amount of shifting of individual sequences.

### A. Finding Helixes

We have chosen the following implementation for our search for variant or consensus helixes. Position two nonoverlapping windows of width W on the data set, at a distance or separation of ℓ apart. Let k be the desired helix length where k ≤ W. Then find the location of the best (if any) helix (or helixes if more than one exists) in each sequence within the specified windows. "Mismatches" correspond to interior loops, while the insertion/deletions of letters correspond to bulges. Usually we will simply search for helixes of some length with a specified amount of mispairing (mismatches). The score for a given window position is the sum of the scores for each sequence. We score a helix by the number of base pairs divided by helix length.

To do a full search of the data for length k helixes in windows of width W, we let ℓ, the separation between windows, vary from ℓ = 0, where there are N − 2W + 1 window

**Table 3**
**ESTIMATES OF STATISTICAL SIGNIFICANCE p FOR APPEARANCE OF CONSENSUS BASE PAIRING (k-LETTER HELICES) BETWEEN TWO WINDOWS OF WIDTH W IN AT LEAST L OF R = 32 SEQUENCES OF LENGTH N = 75**

| k | L | W | α | β | H(α,β) | p |
|---|---|---|---|---|---|---|
| 4 | 16 | 8 | 0.010 | 0.500 | 0.521 | $1.29 \times 10^{-4}$ |
| 5 | 10 | 6 | 0.035 | 0.313 | 0.450 | $1.21 \times 10^{-3}$ |
| 6 | 16 | 25 | 0.010 | 0.500 | 0.521 | $7.23 \times 10^{-3}$ |
| 7 | 8 | 25 | 0.002 | 0.250 | 0.408 | $2.71 \times 10^{-3}$ |

positions, to $\ell = N - 2W$ where exactly one window position is possible. This causes any such search to take $O(N^2)$ time, where W and k are fixed.

### B. Statistical Significance

Once again, it is natural to ask about the significance level of found base pairing patterns. Fortunately the work of the last section can easily be carried over.

The large deviation formulas for statistical significance go as follows. We have R sequences of length N, with a window width W and word size k. The probability α of finding a k letter helix in a given sequence with fixed window positions is

$$\alpha \approx (W - k + 1)^2(1/4)^k$$

since there are $(W - k + 1)$ distinct ways to find the helix location in each window. As above, we want to find a helix at least L of the R sequences. If $\beta = L/R$ and $\beta > \alpha$, the estimate of statistical significance is

$$p = \frac{(N - W)(N - (W + 1))}{2} \exp\{-RH(\alpha, \beta)\}$$

The coefficient of exp{}, $\binom{N}{2}\binom{W}{2}$, in the above equation counts window positions in sequences of length N. For our data set, Table 3 gives some relevant estimates.

Consensus alignment of many RNA sequences is, in principle, a simple straightforward procedure at this point. However, writing usable computer programs is a major problem and, in addition, there are many difficulties encountered in the analysis of actual sequences. Therefore, we illustrate this analysis by folding our set of tRNA sequences.

## IV. FOLDING tRNAs

As emphasized in earlier sections, alignment and folding are interrelated problems. To fold a set of tRNAs, we need a strategy for approaching the problem. Not only are there dependencies between alignment and folding, but also dependencies within both operations of alignment and folding. In folding, for example, there are conflicts between different length helixes as well as quality (number of bulges and interior loops) of helixes. We approach these problems by first locating long (statistically significant) matches between the sequences. Then we locate common patterns of base pairing.

### A. tRNA Alignment by Matches

We first explore alignment by matches. Recall from Section II that while a 4-mer is common to all 32 of our sequences (W = 75), the expected length of pattern common to



FIGURE 3. The six letter pattern gttcga is common to 26 of 32 *E. coli* tRNA sequences. It is the most frequent six letter word.

all 32 was ≈ 2.6. Therefore, without decreasing the window size, we should not consider any pattern less than four letters long. What pattern length k should we begin with? Our approach is to start with larger k and work down to smaller k. The results for k = 6 appear in Figure 3 where gttcga appears in 26 of the 32 sequences. The sequences generally appear as upper case while the patterns we locate appear in lower case. The sequences are right justified in order to highlight these found matches, and the following results support such an alignment. With k = 5, the most common word is ggttc, which is found in 29 sequences and overlaps gttcga in every location where the latter sequence occurs (see Figure 4). Additionally ggttc occurs once upstream (5' or left) of its "canonical" location. In Figure 5, the results for k = 4 are displayed and ggtc is found 32 times, perfectly aligned when the sequences are right justified. Even for k = 3, shown in Figure 6, the 3-mers which are common to all 32 sequences include only subpatterns of ggttcg and cca. Each of these patterns are included in the invariant "positions" of Figure 1, ggttcg(a) being in the TψC loop and cca being the acceptor arm pattern.

Of course, the positions in Figure 1 are the result of an alignment and provide a template for future investigators to fit tRNA sequences to. Here we have set ourselves the task of producing alignment and folding without a template with which to align or fold. Using matches we have now aligned the sequences on what is actually the known TψC loop. Now we turn to finding common folding patterns.

### B. tRNA Alignment by Base Pairing

In Section III, we gave a general description of the procedures to be employed in this section. Here we must present output from our program (fold), and we are required to be

FIGURE 4. The five letter pattern ggttc is common to 29 of 32 *E. coli* tRNA sequences. It is the most frequent five letter word.

specific about some details of our computer method. In particular we will first describe how we organize the (approximately) $N-2W$ graphs of score vs. location of right-hand windows for the (approximately) $N-2W$ separations of the windows. There are about $(N-2W)^2/2$ window positions. Initial alignment of the sequences is also considered. Then we illustrate how base pairings found by the program are displayed along with their connections with the associated score graphs. Then we turn to our analysis of tRNA folding.

### *1. The Score Graphs*

To make matters specific, take the window size, W = 10, and the helix length, k = 5, with the maximum number of allowed mispairs, mm = 0. There are approximately $(N-2W)^2/2$ positions for the two windows. To organize ourselves, we take the horizontal axis to be the position of the rightmost base in the right window for a fixed separation of windows. The vertical axis is score. For the analysis in Figure 7, the sequences are left justified. Each separation is an individual graph. In Figure 7A, all $(N-2W)$ graphs are superimposed, making a jumbled graph. Figure 7B gives a three-dimensional representation of the data. In Figure 7C, an individual graph is given for window separation 3. This peak corresponds to the anticodon stem and is further explored in Figure 9. To relate the single graph for a single separation to all separations, the data in Figure 7A is "redrawn" in Figure 7D keeping the separation = 3 graph solid and plotting all other separations in dotted lines. This procedure allows us to find our way among these complex data.

### *2. Sequence Alignments*

Several sensible alignments of the sequence set are possible. We have already mentioned



FIGURE 5. The pattern gttc is found in all 32 *E. coli* tRNA sequences. It occurs perfectly aligned as well as some other locations.

right justification, which locates the TψC loop. The left justification of Figure 7 locates the anticodon stem. *A priori*, without the probability calculations justifying alignment on TψC, they are equally reasonable alignments. To see that the data analysis differs for left- and right-justified alignments, the superimposed graphs are given for right-justified sequences in Figure 8A and left-justified sequences in Figure 8B.

In all of Figure 8, W = 10, word size = 5, and the amount of mispairing is mm = 0. There is another reasonable possibility for initial alignment; the sequences can be aligned on *both* ends. This simply means that variable loop sizes will result. The superimposed graphs for such a search is presented in Figure 8C.

### *3. Presentation of Helixes*

It is clearly possible, by moving the dotted vertical line, to move about in a graph of a single separation. To move from separation to separation, we move the relative positions of windows on the screen where the sequence set is displayed. To illustrate, Figure 9A is the graph for separation 3, where the sequences are aligned left and W = 10, k = 5, and mm = 0. Corresponding to the horizontal location of the dotted line are the window locations and displayed found pattern of Figure 9B. In summary: (1) moving the right window about in Figure 9B moves the dotted line in Figure 9A and, when separation between windows is changed, moves to another separation graph; and (2) moving the dotted line in Figure 9A moves the window positions in Figure 9B, while maintaining window separation.

Because the TψC loop matches we found above align perfectly when the sequences are right justified, we right justify and scan with W = 10, k = 7, and mm = 0. A highly significant 13 helixes are found in the 32 sequences. In examining this pattern, it is discovered

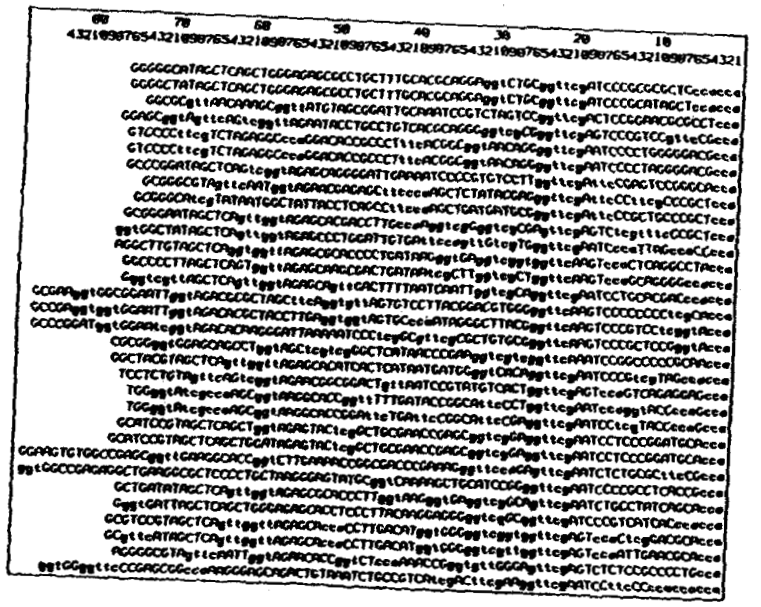


FIGURE 6. Several three letter patterns occur in 32 of 32 *E. coli* tRNA sequences. They are ggt, gtt, ttc, tcg, and cca.

that both right and left justification produces seven letter base pairing in all but three of the sequences! See Figure 10A for a display of this pattern. Notice that in several (seven) places there are actually eight letter base pairings. How is this abundance of base pairings to be handled? If the letters adjacent to the helix are random, then the helix is expected to extend in $1/4 \times 32 = 8$ cases. Therefore we decide not to extend the consensus helix to 8 bp.

Our idea to resolve these difficulties is that of consensus: locate the common features. This removes the ambiguity in all but one sequence, the 11th, which has patterns

ggtggcta ———————— tagccacc.

In this case if

ggtggct ———————— agccacc

is chosen, there is consensus of the left-hand pattern with the other left-hand patterns, but not consensus of the right-hand pattern with the other right-hand patterns. Similarly with gtggcta ———— tagccac, some additional examination is required. If ggtggct ——— agccacc is chosen, this will be the only sequence without a T in the column just 3′ right of the left-hand pattern and, as we will see, the TψC stem will be spoiled. Thus, we resolve the difficulty as in Figure 10B. Allowing one mispairing (mm = 1), we add the three other sequences to the consensus and align on the base pairing in Figure 10C.

The helix located in Figure 10C is, of course, the acceptor stem, involving areas of sequence at the 5′ and 3′ ends. The method of representing helixes by parentheses will



allow us to unambiguously present secondary structure. The cloverleaf of Figure 1 has the symbolic form

5′ ( ( ) ( ) ( ) ) 3′.

In our sequences "( . . . )" and "[ . . . ]" are used to show helix size and location. This scheme, of course, does not work if we do not have secondary structure.

There is no significant pairing with k = 6 when we scan the area between the base-paired regions with W = 10 and mm = 0. Moving to k = 5, we show in Figure 11A the scan with separation 1. The rightmost slender peak corresponds to the TψC stem and is shown (mm = 1) in Figure 11B. The leftmost peak is refined by left justifying the remaining sequences. The consensus pairing pattern is shown in Figure 11C. We, of course, have located the anticodon stem.

Finally, we restrict attention to the segments of sequence between the left-hand k = 7 pattern and the anticodon stem. (Observe the position of the carets, ">" and "<", in Figure 12B). This scan has W = 5, k = 4, and mm = 0. Figure 12A shows one separation (of 7) in dark while the remaining separations are plotted lighter. The pattern corresponding to the peak of the dark line is shown in Figure 12B, and the consensus pattern (mm = 1) is shown in Figure 12C.

This is the complete study of secondary structure for this set of tRNAs, agreeing in detail with that published.[14] Figure 12C is our consensus folding of these tRNA sequences. This is the first time such a task has been accomplished in a mathematically rigorous fashion.

## V. TERTIARY INTERACTIONS

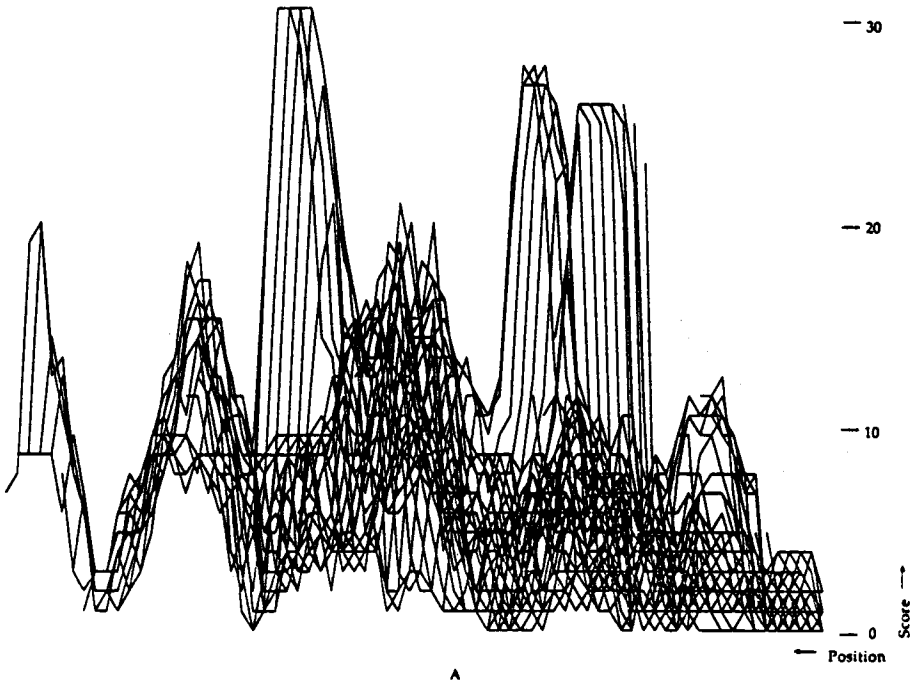
The hydrogen bonding involved in the tRNA cloverleaf is known as secondary structure. See Chapter 7 for a mathematical definition of secondary structure.[1] Viewing the cloverleaf bonds as fixed, additional hydrogen bonds are formed between bases unpaired in the cloverleaf. These additional bonds form what is known as tertiary structure. Figure 13 is a diagram of secondary and tertiary interactions in yeast phe tRNA.[15] The tertiary bonds further fold tRNA into the familiar L-structure found by Kim et al.[5]

Recall that the tertiary interactions are frequently simply additional base pairings and that no changes in pairing rules need to be made for the search. Real difficulty, however, comes with these pattern searches. The sequences are locked into a fairly rigid alignment (see Figure 12C), but no longer is a helix of $k \geq 4$ the object of interest. Instead, Figure 13 shows pairing between single letters. Due to the amount of conserved positions, there is a good deal of potential tertiary interaction. The good news is that such searches are possible; the bad news is that, unlike secondary structure, many conflicting possibilities exist.

Our goal is not to produce a complete analysis of tertiary interactions in tRNA, but to show what is possible with the program and methodology we have presented in this chapter.

Figure 12C shows the cloverleaf produced by our methods. We will now search this alignment for potential tertiary interactions. The D loop and the extra arm are of variable length, and the windows are set (for these runs) to have left and right justification. This should make clear just what alignment is used when the windows are in specified positions.

The first quite naive search is for windows of width 4 and helix size 4 and mm = 1. Thus, no shifting is allowed, even in the variable length regions. Figure 14A shows the full scan, with all separations superimposed. The four collections of peaks which reach maximum value correspond, obviously, to the four helixes of the cloverleaf. Figure 14B shows three of the peaks for a separation of 9 bases. To show some additional results, the medium height set of peaks of Figure 14A, adjacent to the TψC stem peaks, result from possible "pairing" between the left half of the TψC stem (positions 29 to 32) and positions 70 to 73 of the




A

FIGURE 7. Consensus folding analysis with window size W = 10, helix size k = 5, and no mispairs allowed. The sequences are right justified. (A) All separation graphs are superimposed, while in (B), an additional dimension makes a three-dimensional graph. (C) The graph for separation 3 appears alone, and in D, the remainder of the graphs appear with dotted lines.

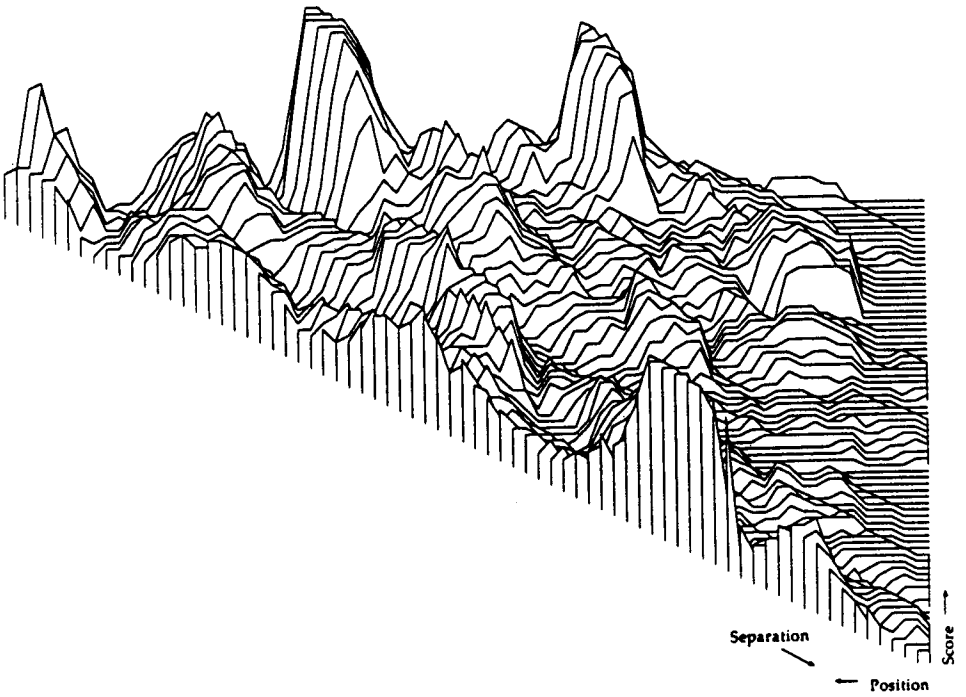



FIGURE 7B.

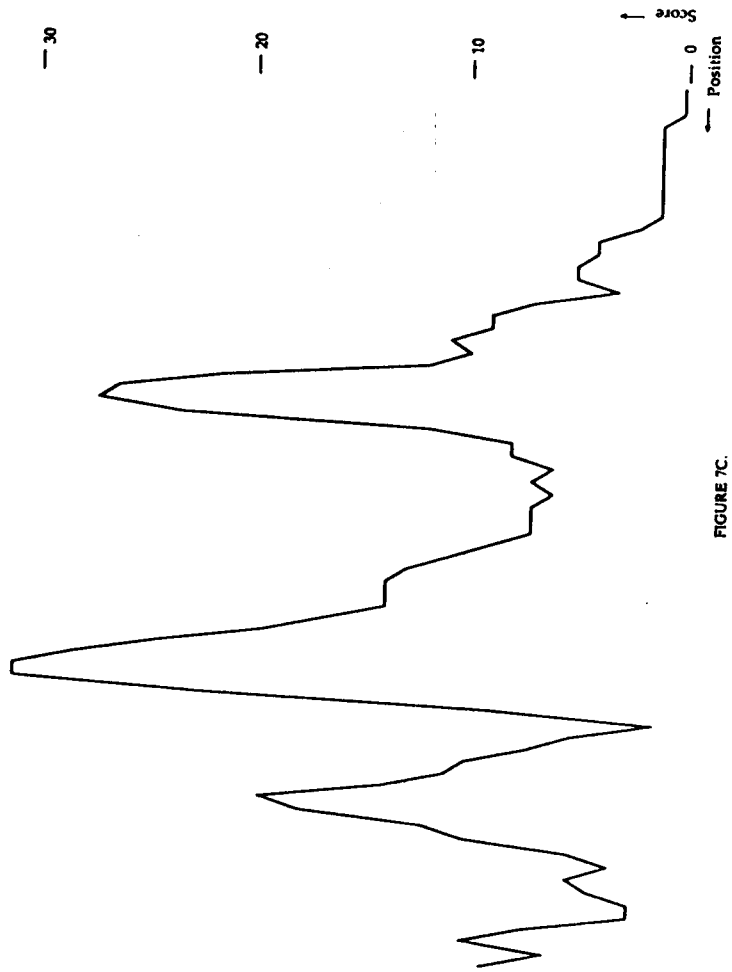



FIGURE 7C.



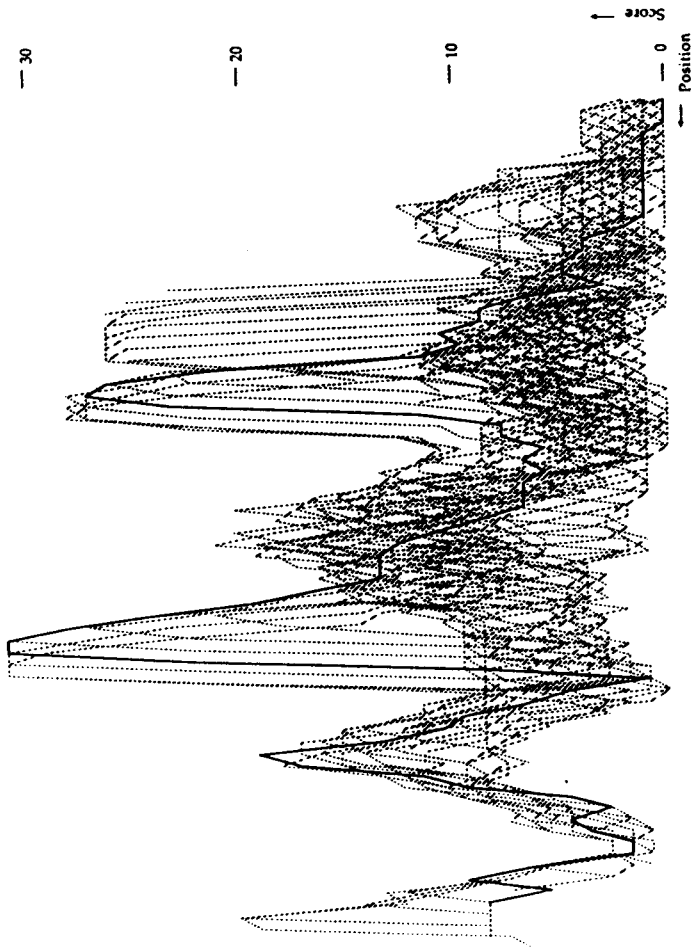
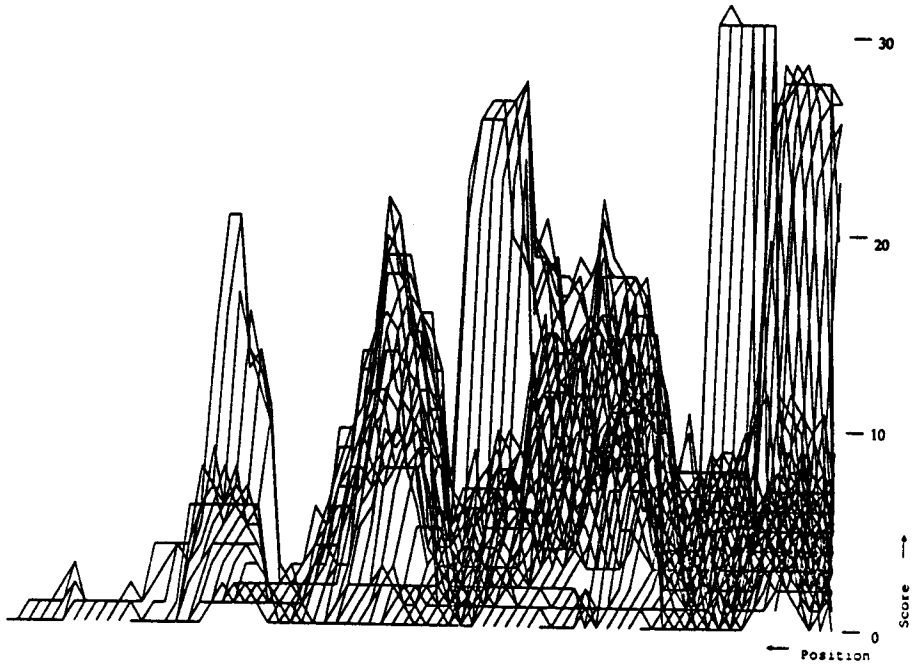



FIGURE 7D.




A

FIGURE 8. Consensus folding analysis with window size = 10, helix size k = 5, and no mispairs allowed. (A) The sequences are right justified; (B) the sequences are left justified; and (C) the sequences are aligned on both ends.

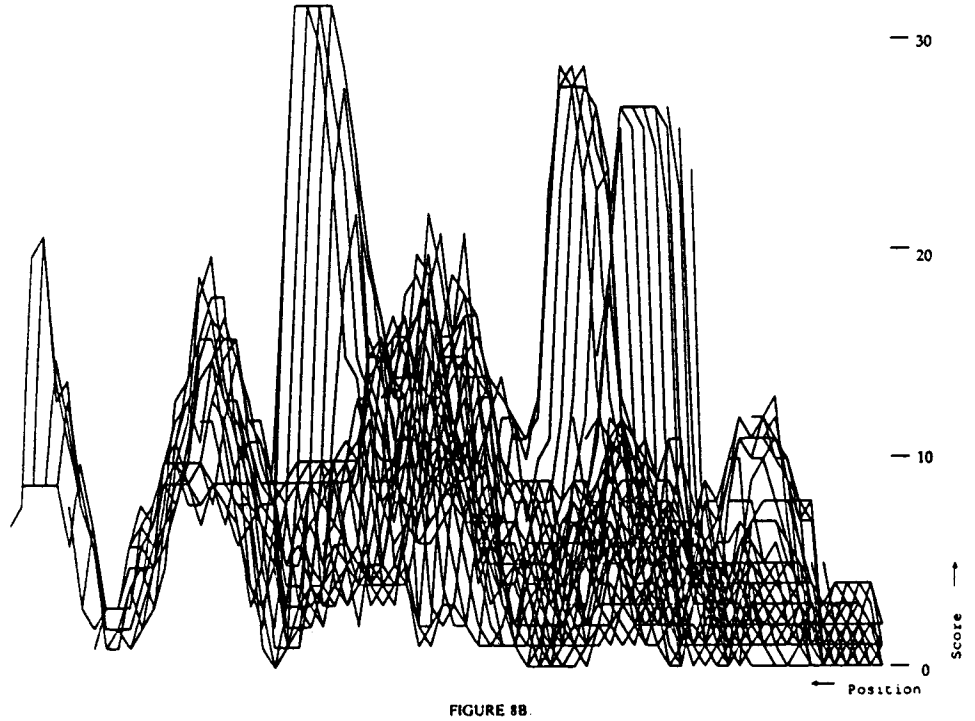



FIGURE 8B.

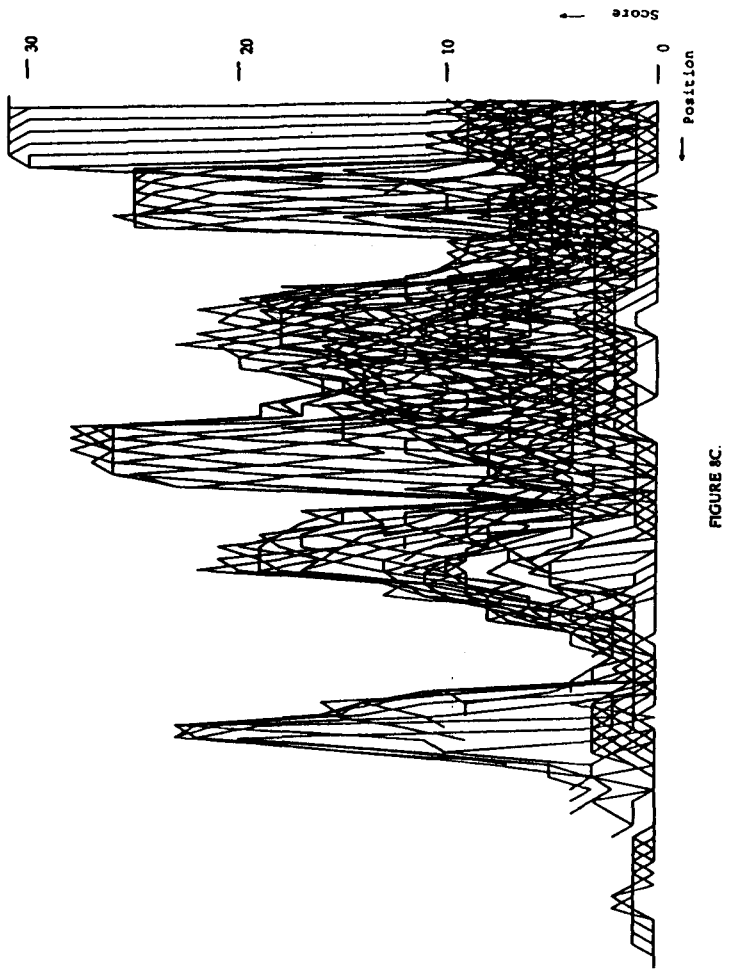
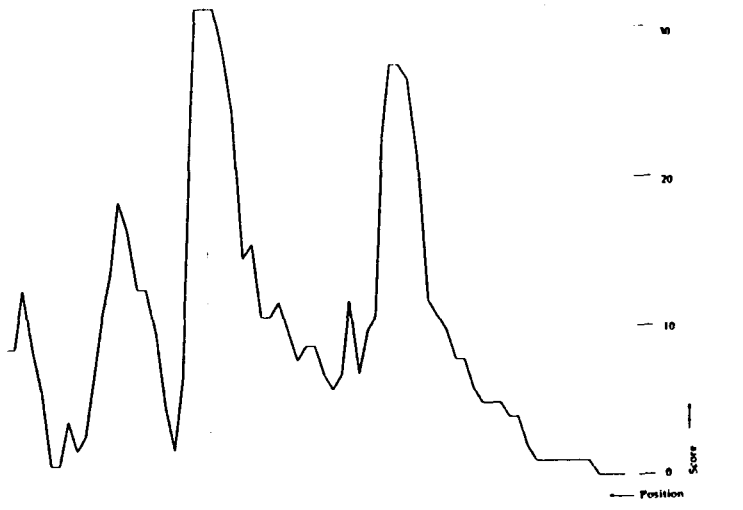
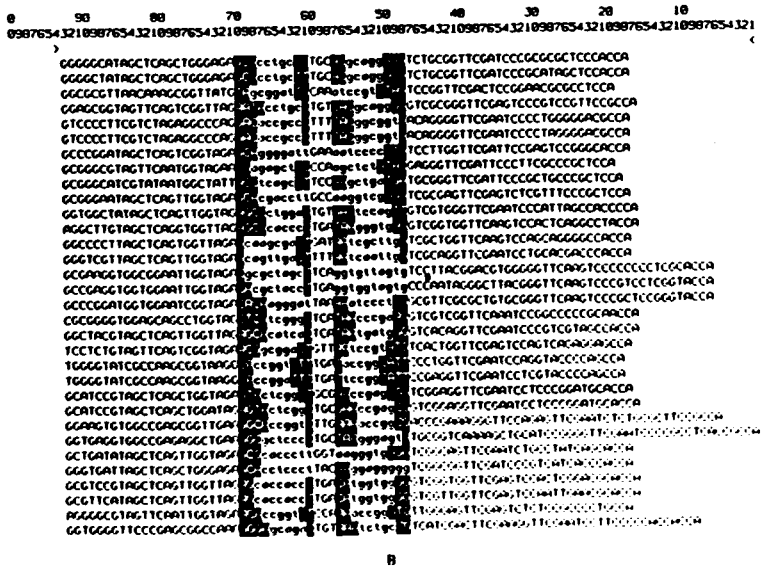



FIGURE 8C.






A



B

FIGURE 9. Consensus folding analysis for left-aligned sequences th W = 10, k = 5, and mm = 0. (A) The graph is for separation 3. The base pairing patterns producing the peak at the dotted line are shown in (B).

A

FIGURE 10. Consensus folding patterns with W = 10, k = 7, and mm = 0. A shows ambiguity in eight sequences. These are removed in B. Allowing 1 mm gives the patterns in C.

FIGURE 10B.

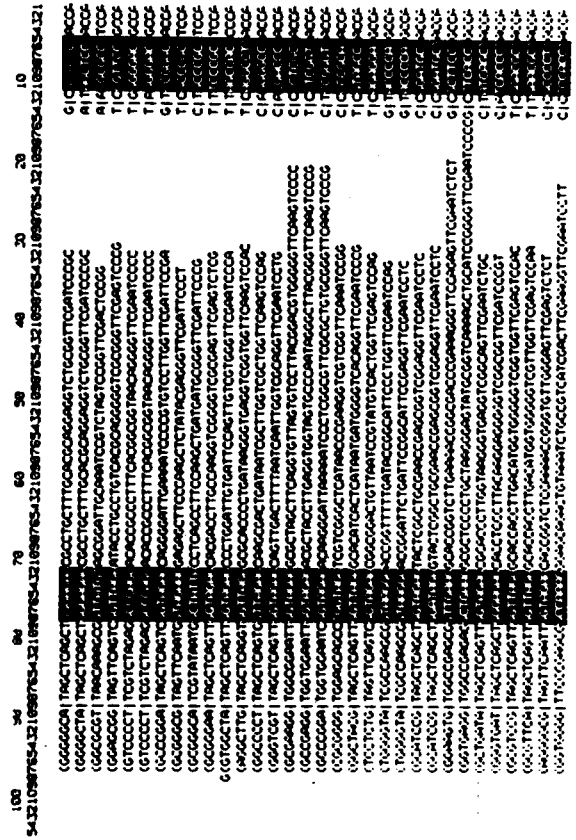


FIGURE 10C.




A

FIGURE 11. Analysis of the configuration of Figure 10C with W = 10, k = 5, and mm = 0. The graph with separation 1 is shown in A and the pattern of the rightmost peak shown in B. The leftmost peak can be refined to the pattern of C. In C, three consensus helixes are represented.

anticodon loop (see Figure 14C for illustration). Now this is not a real interaction, but one which the program easily locates.

We continue our study by showing graphs in Figure 15 of all separations superimposed for two letter (Figure 15A) and one letter (Figure 15B) interactions with no shifting. The strong two letter potential interactions are located as follows:

| Peak | Locations |
|---|---|
| $A_1$: | D stem |
| $A_2$: | 95-96 and 70-71 |
| $A_3$: | Anticodon |
| $A_4$: | 72-73 and 29-30 |
| $A_5$: | 29-30 and 26-27 |
| | 95-96 and 26-27 |
| $A_6$: | 26-27 and 23-24 |
| $A_7$: | T$\psi$C stem |
| $A_8$: | Acceptor stem |

Figure 15B contains most of the actual tertiary interactions. Obviously there is a good deal of data, and both computation and biology are needed to sort out such a situation if the answer is not already understood. It is our hope that computation can prove truly useful in a similar situation where the structure is not known.

FIGURE 11B.

FIGURE 11C.

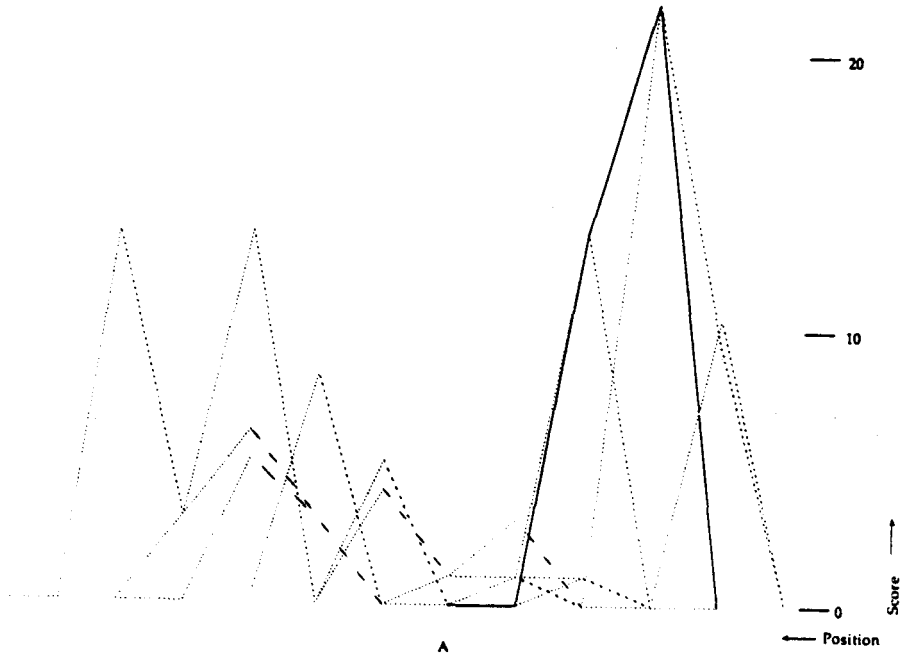



FIGURE 12. Analysis between the left k = 7 pattern and the anticodon stem. Here W = 5, k = 4, and mm = 0. The dark graph of A is for separation 7. The peak has the pattern shown in B which with 1 mm is given in C. C is the consensus secondary structure of these tRNA sequences.

FIGURE 12B.

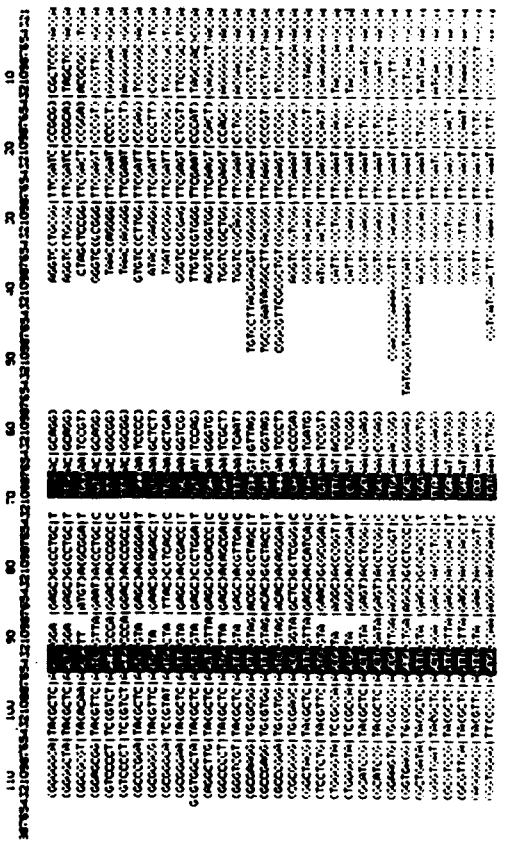


FIGURE 12C.



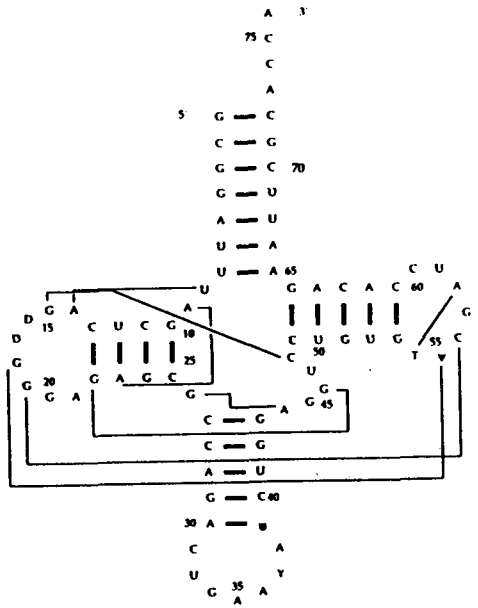


FIGURE 13. Secondary and tertiary structure of yeast phe tRNA.[11]

## VI. CONCLUSIONS

It would seem from the experiences reported in this chapter that the prospects for consensus folding are good, although tertiary interactions might be much more difficult to determine than secondary interactions. Since Levitt's 1969 paper[12] laid the basis that allows this chapter's methods to succeed, it might be asked why the computer development lagged 15 years behind. The reasons relate, it seems to us, to the type of computing previously available: centralized, batch-oriented computer centers. With that resource, it is almost inevitable that the dynamic programming methods be developed first. Dynamic programming is computationally intensive and does not require any human intervention with its recursive calculations. On the other hand, the consensus methods only make sense when some meaningful visual display is provided. Few of us would examine tables of output to recognize where in each sequence a signal was located. Since both reasonable and unreasonable possibilities are produced by consensus, the human is an important part of analysis. Methods such as these are needed to make computational methods into a useful tool for biology.

In the work of Noller and Woese there is the concept of "proven helixes".[20] A helix is said to be proven if there is some base pair of the helix that is distinct from the others in the other sequences. Since these authors are studying distinct organisms, they use the double mutation required to maintain the base pair as evidence for the helix as a real structure. While we have not used this device in the program described here, it is quite easy to include this or other modifications in helix definition or scoring.

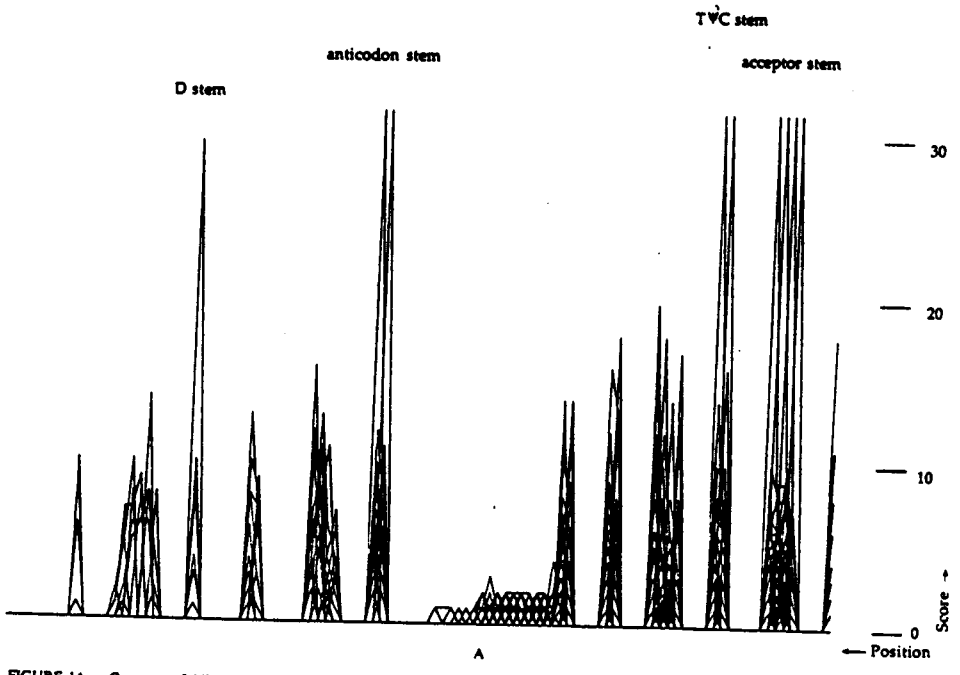



FIGURE 14. Consensus folding analysis of Figure 12C with W = k = 4 and mm = 1. A shows all separations with the four secondary structure helices indicated. B is the graph for a separation of 9 bases. C shows the potential interaction between the left half of the TψC stem and the anticodon loop.

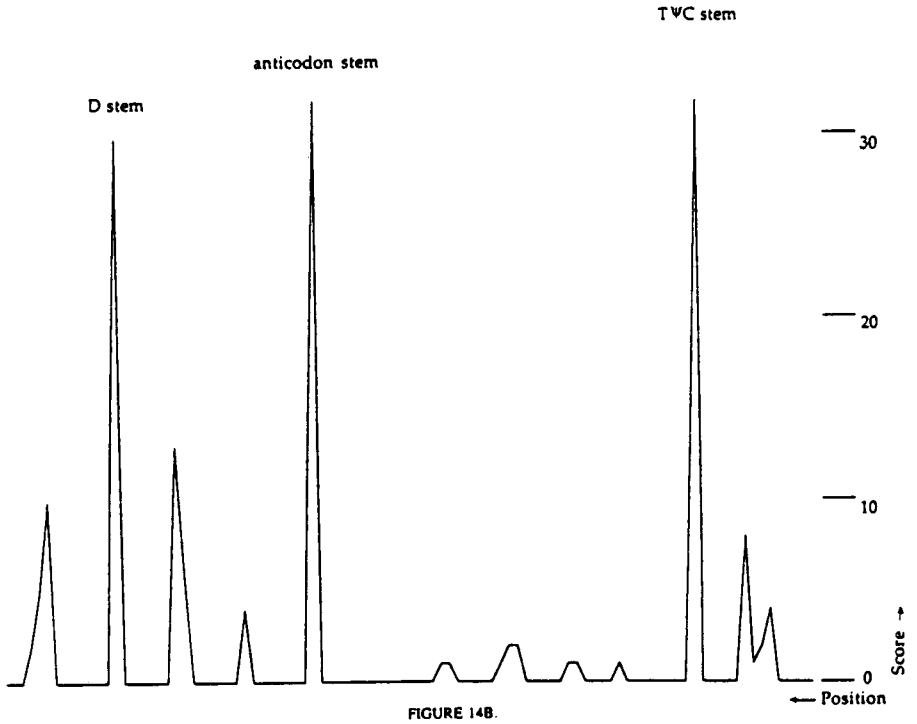
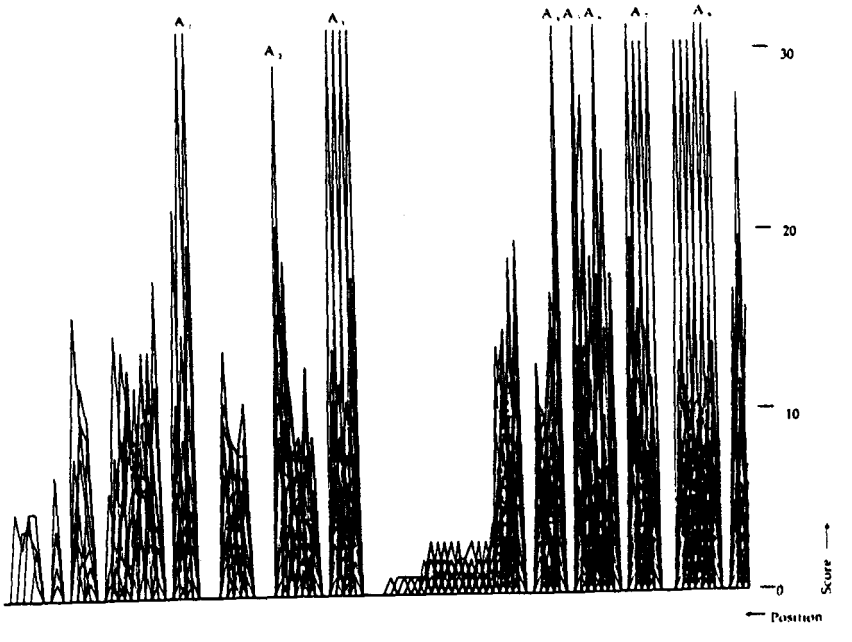



FIGURE 14B.

FIGURE 14C.




A

FIGURE 15. Consensus folding analysis of Figure 12C with W = k = 2 in A and W = k = 1 in B.

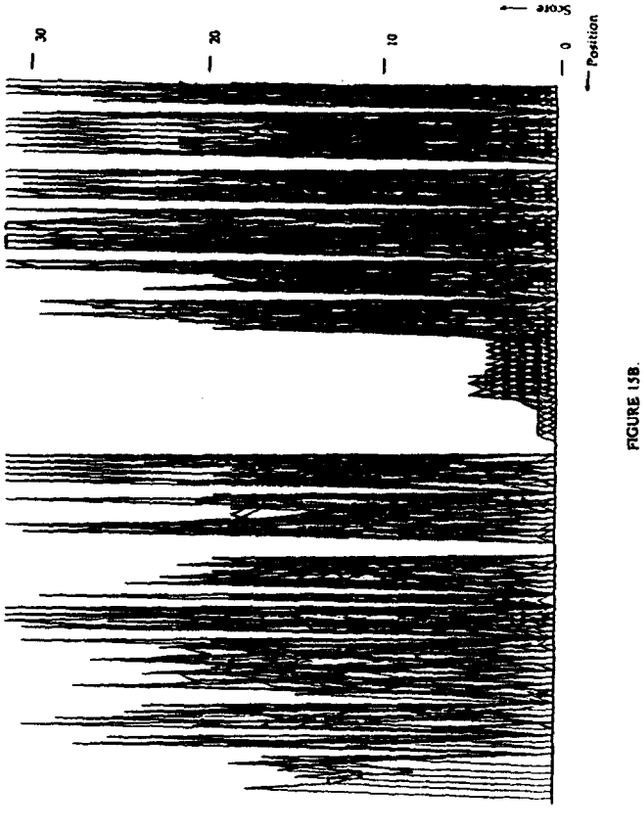



FIGURE 13B.



In closing, we mention that many data sets remain to be examined. The rRNA sequences, 5S, 16S, and 23S, are well analyzed, but it will be instructive and, we hope, revealing to analyze them by these methods. Deeper mathematical and biological questions of inferring rRNA phylogeny via these consensus alignments remain for further study. See Pace et al.[51] for a recent overview of these and related questions.

## ACKNOWLEDGMENT

This work was supported by grants from the System Development Foundation and the National Institutes of Health.

## REFERENCES

1. **Zuker, M. S.,** The use of dynamic programming algorithms in RNA secondary structure prediction, in *Mathematical Methods for DNA Sequences*, Waterman, M. S., Ed., CRC Press, Boca Raton, Fla., 1988.
2. **White, J.,** An introduction to the geometry and topology of DNA structure, in *Mathematical Methods for DNA Sequences*, Waterman, M. S., Ed., CRC Press, Boca Raton, Fla., 1988.
3. **Benham, C.,** Mechanics and equilibria of supercoiled DNA, in *Mathematical Methods for DNA Sequences*, Waterman, M. S., Ed., CRC Press, Boca Raton, Fla., 1988.
4. **Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and Zamir, A.,** Structure of a ribonucleic acid, *Science*, 147, 1462, 1965.
5. **Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H. J., Seeman, N. C., and Rich, A.,** Three-dimensional tertiary structure of yeast phenylalanine transfer RNA, *Science*, 185, 435, 1974.
6. **Tinoco, I., Uhlenbeck, O. C., and Levine, M. D.,** Estimation of secondary structure in ribonucleic acids, *Nature (London)*, 230, 362, 1971.
7. **Stein, P. R. and Waterman, M. S.,** On some new sequences generalizing the Catalan and Motzkin numbers, *Discrete Math.*, 26, 261, 1978.
8. **Waterman, M. S.,** Secondary structure of single-stranded nucleic acids, *Stud. Found. Combinatorics Adv. Math. Suppl. Stud.*, 1, 167, 1978.
9. **Waterman, M. S. and Smith, T. F.,** RNA secondary structure: a complete mathematical analysis, *Math. Biosci.*, 42, 257, 1978.
10. **Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J.,** Algorithms for loop matchings, *SIAM J. Appl. Math.*, 35, 68, 1978.
11. **Sankoff, D.,** Simultaneous solution to the RNA folding, alignment and protosequence problem, *SIAM J. Appl. Math.*,
12. **Levitt, M.,** Detailed molecular model for transfer ribonucleic acid, *Nature (London)*, 224, 759, 1969.
13. **Lewin, B.,** *Genes*, 2nd ed., John Wiley & Sons, Toronto, 1985.
14. **Sprinzl, M., Moll, J., Meissner, F., and Hartmann, T.,** Compilation of tRNA sequences, *Nucl. Acids Res.*, 132, r1, 1985.
15. **Fox, G. E. and Woese, C. R.,** 5S RNA secondary structure, *Nature (London)*, 256, 505, 1975.
16. **Nishikawa, K. and Takemura, S.,** Structure and function of 5S ribosomal ribonucleic acid from *Torulopsis utilis*, *J. Biochem.*, 76, 935, 1974.
17. **Waterman, M. S.,** Computer analysis of nucleic acid sequences, *Methods of Enzymology*, in press.
18. **Trifonov, E. N. and Bolshoi, G.,** Open and closed 5S ribosomal RNA, the only two universal structures encoded in the nucleotide sequences, *J. Mol. Biol.*, 169, 1, 1983.
19. **Woese, C. R., Magrum, L. J., Gupta, R., Siegel, R. B., Stahl, D. A., Kop, J., Crawford, N., Brosius, J., Gutell, R. R., Hogan, J. J., and Noller, H. F.,** Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence, *Nucl. Acids Res.*, 8, 2275, 1980.
20. **Noller, H. F. and Woese, C. R.,** Secondary structure of 16S ribosomal RNA, *Science*, 212, 403, 1981.

21. Stiegler, P., Carbon, P., Zuker, M., Ebel, J. P., and Ebrenmann, C., Structural organization of the 16S ribosomal RNA from *E. coli* topography and secondary structure, *Nucl. Acids Res.*, 9, 2153, 1981.
22. Zwieb, C., Glotz, C., and Brimacombe, R., Secondary structure comparisons between small unit ribosomal RNA molecules from six different species, *Nucl. Acids Res.*, 9, 3621, 1981.
23. Glotz, C., Zwieb, C., Brimacombe, R., Edwards, K., and Kossel, H., Secondary structure of the large subunit ribosomal RNA from *Escherichia coli*, *Zea mays* chloroplast, and human and mouse mitochondrial ribosomes, *Nucl. Acids Res.*, 9, 3287, 1981.
24. Moazed, D., Stern, S., and Noller, H., Rapid chemical probing of conformation in 16S ribosomal RNA and 30S ribosomal subunits using primer extension, *J. Mol. Biol.*, 187, 399, 1986.
25. Noller, H. F., Kop, J., Wheaton, V., Brosius, J., Gutell, R. R., Kopylov, A. M., Dohme, F., Herr, W., Stahl, D. A., Gupta, R., and Woese, C., Secondary structure model for 23S ribosomal RNA, *Nucl. Acids Res.*, 9, 6167, 1980.
26. Karlin, S., Ost, F., and Blaisdell, B. E., Patterns in DNA and amino acid sequences and their statistical significance, in *Mathematical Methods for DNA Sequences*, Waterman, M. S., Ed., CRC Press, Boca Raton, Fla., 1988.
27. Waterman, M. S., Sequence alignments, in *Mathematical Methods for DNA Sequences*, Waterman, M. S., *Ed.*, CRC Press, Boca Raton, Fla., 1988.
28. Martinez, H. M., An efficient method for finding repeats in molecular sequences, *Nucl. Acids Res.*, 11, 4629, 1983.
29. Galas, D., Eggert, M., and Waterman, M. S., Rigorous pattern recognition methods for DNA sequences, *J. Mol. Biol.*, 186, 117, 1985.
30. Bahadur, R. R., *Some Limit Theorems in Statistics*, Society for Industrial and Applied Mathematics, Philadelphia, 1971.
31. Ellis, R. S., *Entropy, Large Deviations, and Statistical Mechanics*, Springer-Verlag, New York, 1985.
32. Pace, N., Olsen, G., and Woese, C., Ribosomal RNA phylogeny and primary lines of evolutionary descent, *Cell*, 45, 325, 1986.