, ⊁⊂`

< 2

. .

- 22. Tremblay, J.-P. and Sorenson, P. G., An Introduction to Data Structures with Applications, McGraw-Hill, New York, 1984.
- Ornateln, R. L. and Fresce, J. R., Correlation of T_u, sequence, and ΔH of complementary RNA helices and comparison with DNA helices, *Biopolymers*, 22, 2001, 1983.
- Waterman, M. S. and Smith, T. F., Rapid dynamic programming algorithms for RNA secondary structure, Adv. Appl. Math., 7(4), 455, 1986.
- Zuker, M. and Stegler, P., Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, *Nucl. Acids Rev.*, 9, 133, 1981.
- Milla, D. R., Kramer, F. R., and Splegelman, S., Complete nucleotide sequence of a replicating RNA molecule, Science, 180, 916, 1973.
- Stlegler, P., Carbon, P., Zuker, M., Ebel, J.-P., and Ebreamann, C., Structural organization of the IdS ribosomal RNA from E. coli. Topography and secondary structure, Nucl. Acids Res., 9, 2153, 1981.
- Vary, C. P. H. and Vournaklu, 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.
- Riemer, D., Colpan, M., Goodman, T. C., Nagel, L., Schumscher, J., Sieger, G., and Hofmann, H., Dynamics and interactions of viroids, J. Biomol. Struct. Dyn., 1, 669, 1983.
- Zuher, M., RNA folding prediction: the continued need for interaction between biologists and mathematicians, Ler. Math. Life Sci., 17, 1966.
- 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.
- Williams, A. L., Jr. and Tinece, I., Jr., A dynamic programming algorithm for finding alternative RNA secondary structures, Nucl. Acids Res., 14(1), 299, 1986.
- Stlegler, P., Carbon, P., Ebel, J.-P., and Ebresmann, C., A general secondary structure model for procaryotic and escaryotic RNAs of the small ribosomal subunits. Eur. J. Biochem., 120, 487, 1981.
- Sanhoff, 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

1.	Intro	duction					
a	Alier	ment b	v Matches	190			
	A Finding Matches						
	<u>л</u> . В	C talia	vicel Significance				
	D.	Junta	The Log(a) Distribution	191			
		1. 2.	The Binomial Distribution and Large Deviations	192			
111	Alia	nment h	w Race Pairing				
ш.							
	А. В.	Stati	stical Significance				
IV.	Fold	ine tRN	IAs				
	<u> </u>	AD N	A Alignment by Base Pairing				
	D.	16.194	The Secre Graphs				
		1.	The Score Orapits				
		2. 3.	Presentation of Helixes				
v	Terr	iary Int	eractions				
•••	C	-					
VI.	Con	Clusion	5	223			
Ack	nowled	gment.	••••••				
Refe	erences						

185

....

I. INTRODUCTION

As the preceding chapter' 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 Band A-forms are both right-handed helixes while the Z-form is left-handed.²³ RNA has basepairing 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 accondary structures. In fact, the first primary (linear) sequence of a tRNA (transfer RNA) appeared in 1965⁴ along with the cloverleaf form of secondary structure. This turned out to be the correct structure and has been verified by X-ray crystallography.⁴ 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 approach we turn to the main topic of this chapter, comparative or consensus analysis of folding.

In an important paper Tinoco et al.⁴ proposed assigning free energies to the components of secondary structure — the various base pairs, end koops, bulges, interior koops, and multibranch koops — 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 1.22×10^{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. Waterman⁹ and Waterman and Smith⁹ 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" has an algorithm to simultaneously fold and align several sequences; three 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, Levitt¹² 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 sequences¹⁴ 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, $(21)^{14} = 3.24 \times 10^{16}$. If R = 32, as in the example of this chapter, $(21)^{32} = 2.05 \times 10^{42}$. If R = 150, a reasonable number of tRNA sequences, $(21)^{39} = 2.15 \times 10^{190}$. In none of these cases is it possible to exhaustively consider all alignments on any modern computer. Allowing gaps from I 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 SS rRNA, which is about 120 bases in length. Fox and Woese¹⁵ and Nishikawa and Takemura¹⁶ 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 Waterman¹⁷ for a consensus approach to folding 34 SS rRNA sequences on a computer.) The difficulties of unequal sequence lengths exist with SS sequences also. A study of Trifonov and Bolshoi¹⁶ studies 5S folding by a related method which has some drawbacks. First, the sequences 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 collaborators²⁴ for 16S rRNA of *Escherichia coli* is shown in Figure 2: These authors describe their approach¹⁹ 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. COLJ (RNA SEQUENCE NAMES WITH THE GenBank ABBREVIATIONS

ala (RNA;	ECOTRAIA
ala (RNA;	ECOTRAIB
cys IRNA;	ECOTRC
asp iRNA;	ECOTRD1
giu tRNA;	ECOTRE1
ghu tRNA;	ECOTRE2
phe tRNA;	ECOTRF
gly (RNA;	ECOTRGI
gly tRNA;	ECOTRG2
gly tRNA;	ECOTRG3
his (RNA;	ECOTRHI
ile IRNA;	ECOTRI
ile tRNA;	ECOTR12
lys (RNA;	ECOTRK
ieu tRNA;	ECOTRLI
iev tRNA;	ECOTRL2
ieu tRNA;	ECTORL5
initiator	ECOTRMF
met tRNA;	
met tRNA;	ECOTRM
ann tRNA;	ECOTRN
gin tRNA;	ECOTRQI
gin aRNA;	ECOTRQ2
arg tRNA;	ECOTRRI
arg tRNA;	ECOTRR2
ser URNA;	ECOTRSI
ser (RNA;	ECOTRS3
the tRNA;	ECOTRTACU
val tRNA;	ECOTRVI
val uRNA;	ECOTRV2A
val tRNA;	ECOTRV2B
tep 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 23S²³ 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 Karlin²⁶ 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(4⁴). If the common sequences are longer, then the techniques of hashing allow the search to be done in NlogN time (see Martinez²ⁿ 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 $N \rightarrow \infty$. 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 X₁, the letters in the sequences, are iid. If the alphabet has four equally likely letters, $p = 4(1/4)^{L-1}$. Let M(R,L) = length of the longest pattern common to at least L of the R sequences. Then,

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

and

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

where

$$\log = \log_{1/p}$$
 and $\gamma \approx .0577$

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

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

= 2.6278 . . .

and

$$\sigma^2 = 0.0842...$$

with

σ = 0.2902 . . .

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 << ∞ . In Figure 5, we could take k = W = 4and 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 kletter sequence,

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

$$= P \begin{pmatrix} W-k+1 \\ U & \{w \text{ starts at } i\} \end{pmatrix} \leq (W-k+1)(1/4)^4$$

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

$$({}^{k})\alpha^{k}(1 - \alpha)^{k-1} = ({}^{k})(W - k + 1)^{k}(1/4)^{k}(1 - (W - k + 1)(1/4)^{k})^{k-1}$$

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

$$({}^{n})(W - k + 1)^{l}(1/4)^{knL-1}(1 - (W - k + 1)(1/4)^{l})^{n-l}$$

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 $\alpha R \ge \beta R = L$ 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 $R = 32$
SFOUENCES OF LENGTH N = 75

k	L	w	a	ß	Η(α,β)	P
4	20	50	0.184	0.625	0.474	1.73 × 10 ⁻³
4	24	75	0.281	0.75	0.472	7.14 × 1013
6	10	75	0.171	0.313	0.662	2.55 × 10.*
7	7	75	0.004	0.188	0.547	4.17 × 10**

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

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 + i possible positions, then the estimate becomes

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

where $\alpha = P(w \text{ occurs in } W \text{ letters})$.

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 7.15×10^{-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 TVC 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 $k \leq W$. 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 $\ell = 0$, 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	•	ß	H(a,\$)	P
4	16		0.010	0.500	0.521	1.29 × 10-4
5	10	6	0.035	0.313	0.450	1.21 × 10-'
6	16	25	0.010	0.500	0.521	7.23 × 10-'
7	8	25	0.002	0.250	0.408	2.71 × 10-1

positions, to l = N - 2W where exactly one window position is possible. This causes any such search to take O(N²) 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

$$\alpha = (W - k + 1)^{2}(1/4)^{4}$$

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 $\{\}, (\overset{w}{i}, \overset{w}{i})$, 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 (RNAs

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

80 70	68	50	40	30	20	18	E 4 1 2 1
176543210907654321090	765432109076	5432109876	543210907	654,32109876	543210987	6543210987	54321

GGGGGCATA	GETICAGETGGGA	CACCECCTEC	TTTGCACGC	ACCACCTCTCC	GgttcgalC		CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCTRTR	CCTCRCCTGGGF	CACCOCCICC	TTICCACCC	ACCACGTCTCC	Gitcgal	CCCCONCCCC	CICC
GGCGCGT	TAACAAAAGCGGT	TATETAGCCC	ATTGCAPAT	CCGTCTAGTCC	GgttcgoCl	CCCC ICCC II	rener
GGAGCGGTAG	TTCAGTCGGTTA	CANTACCTCC	CTGTCACCC	ACCCCGTCCCC	Ggttcgoul		
GTCCCCTTC	GICTAGAGGCCC	CAGGACACCGC	CETTICACO	CCCCTANCACO	agt (cgoA)		er cr cr
GTCCCCTTC	GTCTAGAGGCCO	CAGGACACCGC	CCTTTCACG	GCGGTARCAG	A g t t c g a H l	CCCC HOUSE	COCC
GCCCGGATA	GCTCAGTCGGT	ACACCACCCC	TTCAAAATC	CCCGIGICCI	Gutegal	CCCT ICCCC	CCTCCP
0000000	TAGTTCRATCG	TAGAACGAGAG	CTTCCCAAG	CICIAIHCGA	Agttegall	CCCC TOCCC	CC TCC
CCCCCCCAT	CGIATAATGGC	TATTACCTCAC	CETTCOARC	CIGHIGHIGCO	Agricgal		CTCC
GCGGGAATA	GCTCACTTGGT	ACAGCACGACC	TICCCAACO	100001000	angt i c gao	CCCOTTOCCT	errr
GGTGGCTATA	ICCTCAGT TGGT	ACACCCCTCC	TIGIGATIC	CHETICICO	Sal Contract		CTACC
AGGCTTGTAG	CTORGETGETT	ACACCOCACCO	CTGATAAGO	G TCHUG TCGG			cracci
GGCCCCTTA	GETCAGTGGTT	ACACCANGCO	CTGATAATO	GCT IGG ICGC	00-110-0	TCCTCCBCCA	CACO
GGGTCGTTP	SCTCAGTTGGT	ACAGCAGTTG	CITTIAN (1061000	CCCTTCOOCT		GCACO
6CGARGETGGCGGARTTGGT	AGACGCGCTAG	CTTCAGGTGT	RETETOCT	HUGHEGIGG	CCCTTCOOC	TOTO TOT TOT	GTOCO
CCCGAGGTGGTGGTGGAATTGGT	INGACACCCTAC	CTTGACGTCG	ING I CCCCA	THUGGETTHU	CCCTTCOOC	TCCCCCTCCCC	CTACO
CCCCGGATGGTGGAATCGGT	ACACACARCA	ATTARAAATCO	EICCCCC		CCCTTC000	torecerere to	CAOCO
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	ACCAGCCTGGT	ACCTOGTOCO	CICATARC			TCCCG TCGTO	CCACC
GGCTACGTAC	CIONCTICCIT	ACAGCACATO	AL ICATAAT			TCCOCTCOCA	CACCO
TCCTCTGT	GTTCPGTC66T	ACAACCOCCC	CIGTIANT		TCatton	TCCBCGIBCCT	CAGCO
TGCGGTA	TCGCCAAGCGGT	AUCCUCCCC	THITCATA			TCCTCGTACC	CAGCO
TGGGGTA	TCGCCAAGCGGT	AUCCORCCCC	ATTCICATI		OC_11cgon	TOTTOCCOGA	TGCACC
GCATCCGT	ACTORCES	ACAGTACTCG	GENGEGARO			TECTECCOGA	TGCACO
GCATCCGTR	GCTCAGCTGGAT	AGAGTACTCG	GETGEGARE			TETETGEGET	tecced
GCARGTGTGTGGCCGAGCGGTT	CANCEC ACCES 1	ICTIGAAAACC	GECCACCCG		College	TCCCCCCTC	ACCOCC
GGTGGCCCGACAGGCTGAAGG	COCTOCCTOCT	AAGGGAG TAT	CCGC 10444			TCTCCCTATC	GCACO
GCTGATAT	ACCTCACTTOCT	ACACCOCACC	CIIGGTARG	CG1CHCG1CCC	Cattoga	CICGICATCA	CCCACO
GGGTGATT	ACCTORCTOCO	ACACCACCTO	CC I TACAAG		tC	TCCACTCCCA	CCCACC
CCGTCCGTA	GCTCAGTTGGTT	ACACCACCAC	CITCACATO		ing tege	TCCARTIGAR	CGCACC
GCGTTCATA	OCTOPOTTOGT	TACAGCACCAC	CITCACATO		COntrago	TCTCTCCCCC	CCTGCC
ACCCCCT	ACTICANTICS	TAGAACACCGG	1010000		All a second	000000	ACCACO
GGTGGGGTTCCCCGAGCG	CCCANACGGAGA	CAGACTGTAAA	TCTGCCGTC	AICCACTICA	+ngilegos		

FIGURE 3. The six letter pattern gatega is common to 26 of 32 E. coli uRNA 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 glicga 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 T#C 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 T&C loop. Now we turn to finding common folding patterns.

B. (RNA 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

20 68 10 28 10 876543218987654321898765432189876543218987654321898765432189876543218987654321898765432189876543218987654321898 GGGGGGATAGCTCAGCTGGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGgtteGATCCCCGCGCCCCCCCCCCCC GCGCCTATACCTCACCTCGCAGAGCCCCCTGCTTTCCACGCAGCACCTCTGCGgttcGATCCCCCCATACCTCCACCA GGCGCG11AADAAAGCGG11ATGTAGCGGATTGCAAA1CCG1C1AGTCCGgtteGACTCCGGAACGCGCC1CCA GCAGCGGTAgtteAGTCGGTTAGAATACCTGCCTGTCACGCAGGGGTCGCGGgtteGAGTCCCGTCCgtteCGCCA GTCCCC1TCGTCTRCAGGCCCAGGACACCGCCCT11CACGGCGGTARCAGGGetteGAPTCCCCTGGGGGACGCCA GTCCCCTTCGTCTAGAGGCCCAGGACACCGCCCTTTCACGGCGGTAACAGGGgttcGAAATCCCCTAGGGGACGCCA GCGGGCATCGTATAATGGCTATTACCTCAGCCTTCCAAGCTGATGATGCGGgtteCATTCCCGCTGCCCCCCCCC GCGGGARTAGCTCAGTTGGTAGAGCACGACGTTGCCAAGGTCGCGAgtteGAGTCTCGTTTGCCCCCCCG GGTGGCTATAGCTCAGTTGGTAGAGCCCTGGATTGTCATTCCAGTTGTCGTGGgtteGAATCCCATTAGCCACCCCA AGGCTTGTAGCTCAGGTGGTTAGAGCGCGCGCCCCGATAACGGTGAGGTCGGTGgttcAAGTCCACTCAGGCCTACCA GCCCCCTTRCCTCACTGGTTACAGCAAGCGACTGATAATCGCTTGGTCGCTGgttcAAGTCCACCAGGGGCCCACC GGGTCGTTAGCTCAGTTGGTAGAGCAGTTGACTTTTAATCAATTGGTCGCAGgtveGAATCCTGCACGACCCACCA GCCCAGGTGGTGGAATTGGTAGACACGCTACCTTGAGGTGGTAGTGCCCAATAGGGCTTACGGgtteAAGTCCCGTCCGCTCCGTACCA GCCCGGATGGTGGAATCGGTAGACAGAGGGATTAAAAATCCCTCGGCgtteGCGCTGTGCGGgtteAAGTCCCGCTCCGGGTACCA GGCTACGTAGGTGGGTTGGTTAGAGCACATCACTCATAATGATGGGGTCACAGgtteGAATCCCGTCGTAGCCACCA TCCTCTGTAgtteAGTCGGTAGAACCGGCGGACTGTTAATCCGTATGTCACTGgtteGAGTCCAGTCAGAGGAGCCA TGGGGTATCGCCAAGCGGTAAGGCACCGGTTTTTGATACCGGCATTCCCTGgtteGAATCCAGGTACCCCAGCCA TGGGGTATCGCCAAGCGGTAAGGCACCGGATTCTGATTCCGGCATTCCGAGgttcGAATCCTCGTACCCCAGCC GCATCEGTAGETGAGTGGTAGAGTACTCGGCTGCGAACEGAGCGGTCGGAGgtteGAATCETECCGGATGCACCA GCATCCGTAGCTCAGCTGGATAGAGTACTCGGCTGCGAACCGAGCGGTCGGAGgtteGAATCCTCCCGGATGCACCA GG TGGCCGACAGGCTGAAGGCGCTCCCCCTGCTAAGGCAGTATGCGGTCAAAAGCCTGCATCCCGGg LLeGAATCCCCGCCTCACCGCC GCTGATATAGCTCAGTTGGTAGAGCGCACCCTTGGTAAGGGTGAGCGTCGGCAgttgGAATCTGCCTATCAGCACCC GGGTGATTAGCTCACCTGGGAGAGCACCTCCCTTAGAAGGGGGGGTCGGCCGgttgGATCGCCCCCCCACCA GCGTCCGTAGCTCAGTTGGTTAGAGCACCACCTTGACATGGTGGGGGTCGGTGgtteGAGTCCACTCGGACGCACCA

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

right justification, which locates the TVC loop. The left justification of Figure 7 locates the anticodon stem. A priori, without the probability calculations justifying alignment on TVC, 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 T&C 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 holixes are found in the 32 sequences. In examining this pattern, it is discovered

196 Mathematical Methods for DNA Sequences

60

70

GEGGGCATAGCTCASCTGGGAGAGCGCCTGCT1TGCACGCAGGAGGTCTGCgg11EGATGCCGCGCGCTCCCACCA debotarine, tore, to account of the consense o CONCCUSTANT TOAGTCOGT TACANTACCTOCCTGTCACCCACCCCGTCCCCGggt tecAngTCCCCGTCCCGTTCCCCCA GTCCCCTTCGTCTAGAGGCCCAGGACACCGCCCTTTCACGGCGGTAACAGGggtteGAATCCCCTGGGGGACGCC GTCCCCTTCGTCTAGAGGCCCAGGACACCGCCCTTTCACGGCGGTAACACGggtteGAATCCCCTAGGGGACGCC GCCCGGATAGCTCAGTCGGTAGAGCAGGGATTGAAAATCCCCGTGTCCTTggttcGATTCCCAGTCCGGGCACCA CCCCCAPITACCTCAGTTGGTAGAGCACGACCTTGCCPAGGTCGCGGGTCGCGAGTTCGAGTCTCGTTTCCCGCTCCA GETEGE TATAGETORE TEGETAGAGECEETGGAT TETCAT TECAST TETESTATUCESTATAGECACTECAT TAGECACECCA AGECTTE TRECTCREETERST TREASOCCACCCC TEATARCEGTERGETCGET 99 1 1-896 TECRCT CREECCTACCA GGGTCGTTAGCTCAGTTGGTAGAGCAGTTGACTTTTAATCAATTGGTCGCAggetteGAATCCTGCACGACCCACCA GGCTRCGTRCCTORGTTGGTTRGRCCACRTCRCTCRTRATGATGGGGTCRCAyetteGATCCCGTCGTRGCCACC TOCTCTGTAGTTCAGTCGGTAGAACGGCGGACTGTTAATCCGTATGTCACTggtteGAGTCCAGTCAGAGGAGCCG TEGESTATCSCOARCESTARCCORCESTTTTTCATACCESCATTCCCT#915CGAATCCASSTACCCCASCCA TEGGETRYCGCCRAGCEGTRACGCRCCGGATYCTGATYCCGGCRTTCCCAggeteGRATCCTCGTACCCCAGCC SCATCESTAGETCAGETCAGTAGAGTACTCCCCTGCGAACCCAGCGGTCCGAggtteGAATCCTCCCGGATGCACC CONTECT THE CONSECUTION OF THE CONTECT OF THE CONTE GETERCUCARGAGECTERARGECECTCCCCTCCTARGEGAECTRIECCEGTCARARGECTECRATCCCCGGgg112CCARTCCCCCCCTCACCGCC SCTGATATAGCTCAGTTGGTAGAGCGCACCCTTGGTAAGGGTGAGGTCGGCAGTTCGAATCTGCCTATORGCACCA CCTARINING, CORE TOO TRANSCORDULT TOO TRANSCORDE TO CORE TO LOANT, THE TRANSCORD TO CORE TO CO COTTONIAC TORTTOSTTOSTCOCCCCTTOROTICS TEGESETCCTTING TORATIONAL TORACCOCC RESECCTACTIONITIES TRANSPORCESS TO TORATION TO THE TORATION TO THE TORACCOCCCTTOCCC

FIGURE 4. The five letter pattern ggnc 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 $(N - 2W)^{2/2}$ 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 smong these complex data.

2. Sequence Alignments

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



FIGURE 6. Several three letter patterns occur in 32 of 32 E. coll sRNA acquences. They are ggi, gti, itc, icg,

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

ggiggcta -----Lancace

In this case if

SEIRECI ----- Ageceace

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 ---- tagocac, some additional examination is required. If ggtggct -agocacc is chosen, this will be the only acquence without a T in the column just 3' right of the left-hand pattern and, as we will see, the T&C 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 S' 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 $k \approx 6$ 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 TVC 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 = 7pattern and the anticodon stem. (Observe the position of the carets, ">" and "<", in Figure (2B). 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." The tertiary bonds further fold tRNA into the familiar L-structure found by Kim et al.³

Recall that the tentiary 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 $k \ge 4$ 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 TAC stem peaks, result from possible "pairing" between the left half of the TVC stem (positions 29 to 32) and positions 70 to 73 of the











FIGURE 9. Consensus folding analysis for left-aligned sequences th $W \simeq 10$, 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).







iη.

. . .

.

121 MS94080121 MS9	610 817 - 100 817 - 100 800 818 84 - 40 100 100 100 100 100 100 100 100 100 1					
242210000				202	3000	riciter Ticeantico
	88.8	3888	- 8285	Serios Se		TIC CCTC MACTIC MACTIC CCA CCA CCA CCA CCA CCA CCA CCA CCA C
	TTCCATCO TTCCATCO 00ACTCCC	TTCCARTCO TTCCARTCO	TCCATTCCC TCCATTCCC TTCCAGTCC TTCCAGTCC	TTCMCTCC TTCCMTCC TTCCMTCC CCCCTTACC	CICCTG TCCC GTTCCAATC GTTCCAATC GTTCCAATC TTCCAATCC TCCAATCC TCCAATCC	TTCCAATC 6TTCCAATC ACCCTTCCAATC ACCCTTCCAATC TTCCAATCTCC TTCCAATCCCC 6TTCCACTCC
	Generace Generace Metoceen	TIMORCOCCC TIMORCOCCCC TGTCCTTCC	carcocol carcococi corcococo retectoco retectoco	ACTOCOMO ACTOCOMO ACTOCOMO ACTOCOMO ACTOCOMO	200000 1000 800 100 100 800 100 100 800 100 100 100 100 100 100 100 100 100 100	controcate controcate carcecte androcate contracte contracte contracte contracte contracte
	ACCACCA ACCCACCA ACCCACCA ATCCCACCA		ACCTCAT ACCTCAT ACCTCCC ATTCCACT	MICCCTT MICCMIT MICCMIT COLOCITIC COLOCITIC	TRACCCCA TRACCCCA TRATCATCA TRACCCCA TRACCCCA TRACCCCA TRACCCCCA	
	CTGCTTTGC CTGCTTTGC GGATTGC00	DCCCCCTTT DCCCCCTTT DCCCCCTTT DCCCCCTTTC		CCACTCAT TCACTTTT TRCCTTCA	CCCCTCA CCCCCCCC ATCACTCA CCCCCCC ATTTTCA CATTTTCA	reactede Treactede Consterta Treacted Constrain Constrain Constrain
		998 				
			ATANC C TOAGTT C TOAGTT C	TCNGTC 1 TCNGTT 2 CCMNTT 2		TCACT I TCACT I CONCCT I CONCCT I CONCCT I CONCCT I TCACT I TCACT I TCACT I TCACT I TCACT I
	CCA THC CTA THC CCA THC CCT THC	CCT TCCT CCT TCCT CCT TCCT	CONTROL	CCT TRCC CCT TRCC CCT TRCC PCC TCCC	1000 1000 1000 1000 1000 1000 1000 100	CCS 1900 CCS 1900 610 1000 810 1000 810 1000 700 1900 700 1900 700 1900 700 1900 700 1901
		5533	22253 22253 22253 22253	9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9		
Ĩ,						



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 helizes 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,:	D stem
A2:	95-96 and 70-71
A,:	Anticodon
A.:	72-73 and 29-30
A.:	29-30 and 26-27
•	95-96 and 26-27
A.:	26-27 and 23-24
A,:	T∳C stem
Α.	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.

		ACCARACTTCCCCCCGGIC ACCARACTTCCCCASCTCCCCGGIC TPACTCASCTCCCASCTCCCASCTCCCCCCGGIC TPACTCASCCTCCCASCTCCCASCTCCCGGIC TPACTCASCCTCCASCTCCATCAT	ACONCRACTTECORRECTE ACONCRACTTECORRECTE ACONCRATTECORTICORTICE ACONCRATTECORTICE ACONCRATTECORTICE ACONCRATTER ACONC	Concrection and a concrection of the concrection of	CTR-CTTG-ACTUTACTGCTCCTACCAACGT (CCCCG) TTC-AMI (CCTCC) (A 25 CTA-CTTG-ACTUCTACCAATACCCCAATACCCCT (CCCCG) (TTC-AMI (CCCCC) (C 272 CTA-CTTG-ACTUCTCAATACCCCAATACCACTT (CCCCC) (C 272	0/ATTAINARTCCCTCccccTct coccc1 Trons Ticonc1 In 2000 0//CTVATAINCCCCARCCCCCTct coccc1 Trons Ticonc2 In 2000	2011011000010 (ACACO 17000010000000000000000000000000000000	ILI MARKAGATTC CCT66 TTC6AAT CC466 TTC6AAT CC466 TTC6AAT TTC6466 TTC6AAT CC766 TTC6AAT TTC7AAT CC766 TTC6AAT CC766 TTC6AAT CC766 TTC6AAT CC766 TTC6AAT CC766 TTC6AAT TTC6AAT TTC76AAT CC766 TTC6AAT CC766 TTC76AAT CC766 TTC76AAT TTC76AAT CC766 TTC76AAT TTC76AAT CC766 TTC76AAT CC766 TTC76AAT CC766 TTC76AAT CC766 TTC76AAT CC766 TTC76AAT CC766 TTC76AAT TTC76AAT CC766 TTC76AAT CC766 TTC76AAT CC766 TTC76AAT CC7		CTACTARAZORIATICCGATCARAGCTCGATCCCCCCGATT COMPT [CTCT45] [C_1100 TOTTARAZORIATICCGATCARAGCTCGATCCCCCCCGAL TTCARATT [CTCC45] [C_1010 TOTTARAZOTTARAZOTC TOTTARAZOTTARAZOTC	CTT
INSTRUCTION CONTRACTOR				CARCICACTEGATICS (CASTGATTORET) (CASTGATTORET) (CAST) CASTGATTATALCASTCASTC CASTTARICASTCASTC	ACTAXCTTGAACTATRACTGTCCTTACCAACCT (4000000000000000000000000000000000000	WATTAMMATCCTCGCGTTCGCGCTGT (GCCGG TTCMGT GCCCC)	24.1041441041000010 (4040011100041) 4.10414410014610 4.10414410014610 (4010401100041)	TOTATING CONCENTS TOTATING CONC	Trochestication (concerning trochestication)	CTACTHARGENETCEGATOANNACTCONTCICCOM TTCANT CTCTG) TOTTO	CTTT-STREAM CONTRACTOR (CONTRACTICOGNE) CTTT-STREAM CONTRACTOR (CONTRACTICOGNE) CTTT-STREAM CONTRACTOR (CONTRACTICOGNE) CTTT-STREAM CONTRACTIC (CONTRACTICOGNE) CTTTCS)
			RECREARCT TECENARIA TOCAR RECORDENCE TOCARGE TOCARGE TOCAR RECORDENCE AND TOCARTIC TOCAR RECORDENCE AND TOCARTICS		CTAXCTTCXXCTTAGTGTCCTTAGCACGT (CCCCC01111CCC CTACCTTCXXCTAGTGTCCCTAACACGCCT (CCCCCG11TCCAAC CTACCTTCAXCATAGTGCCCCAATAGCACTT (ACCCG11TTCAAC	0/4/11/4/441CCCTC66CGTTC6CCCTGT (6C266/11/2006 0/0/211/4/14/CCC646CGTC6CCCTGT (6C266/11/2046	ANTERTATCOCATC (ACACC) TTCCAR ATTTTCCTATCCC TTTCCAR ACTCC TCCAR ACTCC TCCAR	TELEVIENCESCATTC CCTCC) TTCCAN TELEVIENCESCATTC CCCCCAN (CCTCC) TTCCAN	TO COMPACE CONTRACTOR COMPACE TO COMPACE TO COMPACE TO COMPACE COMPACE COMPACE COMPACE COMPACE TO COMPACE	CTVCTNAPCCARTATCCGTCOARAGCTCCATCCCCCCG TTCCAR T001746/X0TCAPCTC T001746/X0TCAPCTC	
TERMENT TO THE TANK T	RGARDACOCCTTTCACCACCTARC (ACC) RGARDACOCCTTTCACCACCGTARC (ACC) CACCACCACTTCAASAATCCCCCTARC		RECREACTTACCARGETCALAGETC RECORDANTESTATTACTACTURE RECORDANTESTATTACTACTURE REALEGRETCATTATACTACTACTACT	CATGACTICATIANTCCTTCCTC CATGACTICATICCTTCCTC CATTCATTCATTCCTC CACTTCATTCATTC	CTMXCTTCMXCTTTectGTCCTTRCCACGT (CCCC CTACCTTCAX6TG6TACTCCCTARTACCACTT (ACCC	004TTM-MARTCCTCGGCGTTCGCCCTGT (GCCC 004CTTM-MACCCCMMGGTC (GTCG	ALTCHTANTCATCCCCTC (ACAC 4.TGTTANTCCCTATCTC 11TGATCCCTATCTC (ACTC	TUTHINGCOCONTC TUTHINGCOCONTC TUTHINGCOCONTC	00.000012/04000010 (00000 21/2/200000010 (00000 31/04/10000000010 (00000	CTVGTHM2024CTATCCGGTC2ADAGCTGCATCCCCCCC T027444567C4ACTACCGGTC2ADAGCTGCATCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	0.000 0.111-0.4100100001C 0.0111-0.41001000001C 0.0110-0.4100100101 0.0100 0.00000 0.0000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.000000 0.00000 0.00000 0.00000 0.00000
TO ACCOUNT VALUE CONTACT	MGACACCCCTTTCACGCGGTAAC MGACACCCCTTTCACGCGGGTAAC GACGGGGGATTGAAAATCCCCGGTAT	ACCACACCTTCCCAACCTCTATAC	Recretence Trecement research Reconstruction Treasers Reconstruction Treasers	Conceastranteers Conceastranteers Conteastranteers	at much toward the test accelerated acceleration of the second states the second states acceleration acceleration of the second states acceler	offer Thinking Contract Contract of Contract of Contract Contract Contract Contract of Con	Jejichtentocooto 4.16.116atocotatoto 11116ato-cootato	TO TO A TO COLORATION TO TO COLORATION TO TO COLORATION	restrationers Strationersers Stationersers	CTACTAACCACTACCCCCCCCCCCCCCCCCCCCCCCCC	00000000000000000000000000000000000000
Interesting and	MGMCMCCCCTTCACGCCGTAAC MGMCMCCCCCTTCACGCCGGTAAC MGCMCGCCCCTTCACGCCCGTAAC	ACCACACCTTCCCAACCTCTATAC	MCCREACETTCCCAACGTCCCCCCCCCCCCCCCCCCCCCCCCCC	CARCCACTCATANTCCCTTCCTC	AT AXETTE A STUTT A STORE CONTACCO	onal Thinkart CCCT CCCCTTCCCC	Jay Tombargargagge 4.7677447CGGTA7GTC 7777647646664467	TUTONICOCCALATIC	restructions Transacconcector Transacconcector	CTVCTMMCCAGTATCCCGTCMD TOTTMACCAGCTCAGCTC	00111
Interface of the second se	MURCHCOSCCTTTCACOCCGTAN MURCHCOSCCTTTCACOCCGTAN MUCHCOSCOCTTTCACOCCCGTAN	PACCACCTTCCCAACCTCTATAC	RECRETERECT RECEMENT CONCRETERE	CARCONCINATION CONTINUES OF CON	ACTACCT CASA TO TRATIC TO ACTACCT TO ASA TO COTACT OCC	orial Transmitteeteece oriest chinaateeceaneer		TCT/ATTCCCCCATTC	Trefferences	CTNCTNNCCACTATCA TONTANNOCTOANCTC	01110-01100000000000000000000000000000
If Charles Contractors Income	MULACACCOCCTTTCACO ACAACACCOCCTTTCACO ACCACOCOCCTTCASOA	PECHOROCTTCCCM-CT	ACCCCACCTTCCCAACC ACCCTCCATTCTCCCAACC	CARCCACTCATANTCC	2014/0110/420101 201400110/420100	CONTRACTOR		PUTATION C		0700764000 700764000 770	
TCTACCCATT INCANTACCTCC	RGRORCCCC RGRORCCCCC AGGRORCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	PCCACACCTTC TTACCTCACCTTC	ACCCCCCCCTT ACCCCTCCATT ACCCCTCCATT	CARCEACT	CTACCT CTACCT	58	ιš	22	265	525	
	558		888	603	***	i i i			0000		
Ē	U U U U	25	CTRC TRC	TTACAG	GTACACCC GTACACCC		THEAD		ATACACT TCAACC	Treased Thereout	Treason
CARCCC TTCACCC	1019000	TTC-ATCC TATAATCC	1046116 1046116 1046116	TCACTCC	000000000000000000000000000000000000000	SCACCCT	TUNCTOR	CHECCO.	TCASCTCC	TCHCHCCC TCHCTTCC TCHCTTCC	
AGT INC AGG INC AGG INC	Seri 1125	2 (lum) 9 (lum)	Sulers Sulers			01100	TG ROT	ang rest contress	0011-00 1011-00	tel 162 Mel 162 Al 1762	8888 8888
888 888 885	58	88		88	3 8 8 9 9 9		Ciccle Circles		88		2 E 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
	1937 - 2000 1937 - 2000 1937 - 2000	11007 200) 11007 200) 11007 200)	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2			39999999999999999999999999999999999999					

GURE 118.

127753UB0127753UB0127753UB0127753UB0127753 01 02 07 07 07	CITCCCCI ITTCCATC ICCCCCI ICCCTCCC ACC CCCCCI ITTCCATC ICCCCCI ICCCTCCC ACC CCCCCI ITTCCATC ICCCCAT I ACCTCC ACC CCCCCI ITTCCATC ICCCCAT I ACCTCC ACC CCCCCI ITTCCATT ICCCCAT I ACCTCC ACC CCCCCI ITTCCATT ICCCCAT I ACCCCATCCC ACC CCCCCI ITTCCATT ICCCCAT I ACCCCATC CCCCCI ITTCCATT ICCCCAT I ACCCCATC CCCCCCI ITTCCATT I ACCCCATC ACC CCCCCI ITTCCATT ICCCCAT I ACCCCATC CCCCCI ITTCCATT I ACCCCATC ACC CCCCCI ITTCCATT I ACCCCATC ACC CCCCCI ITTCCATT I ACCCCAT I ACCCCATC CCCCCCI ITTCCATT I ACCCCATC ACC CCCCCI ITTCCATT I ACCCCATC ACC CCCCCCI ITTCCATT I ACCCCATC ACC CCCCCI ITTCCATT I ACCCCAT I ACCCCATC CCCCCCI ITTCCATT I ACCCCATC ACC CCCCCCI ITTCCATT I ACCCCATC ACC CCCCCI ITTCCATT I ACCCCATC ACC CCCCCCI ITTCCATT I ACCCCATC ACCCATC CCCCCCI I TTCCAATT I ACCCCCATC ACCCATC CCCCCCI I TTCCAATT I ACCCCATC ACCCATC CCCCCCI I TTCCAATT I ACCCCCATC ACCCA CCCCCCI I TTCCAATT I ACCCCATC ACCCATC CCCCCCI I TTCCAATT I ACCCCCATC ACCCA CCCCCCI I TTCCAATT I CCCCCCI I ACCCCCATCCAACCAA CCCCCCI I TTCCAATT I CCCCCATC ACCCAACCAACCAACCA CCCCCCI I TTCCAATT I CCCCCI I ACCCCCAACCAACCAACCA CCCCCCI I TTCCAATT I CCCCCI I ACCCCCAACCAACCAACCAACCAACCAA	
100121103816 100801	2000 200 2000 2	RE LIC.
10121753/360121		FIGU
80 1754 121 0987654	ALARCAS (CCTOC) ALARCAS (CCTOC) TREAMING (CCTOC) TREAMING (CCTOC) TREAMING (CCCOC) CORRECTION CORRECTION CORRECTION ALARCAS (CCCOC) ALARCAS (CCCOC) A	
180 36 216387654.121638		
110 110 110	0000000 0000000 0000000 0000000 000000	

212 Mathematical Methods for DNA Sequences









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 paper¹² 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".³⁰ 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.







0 ន

= 2 in A and W = k = 1 in B.

ž

FIGURE 15.

<



In closing, we mention that many data sets remain to be examined. The rRNA sequences, 55, 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.¹² 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

- 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, Fia., 1988.
- 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, Fis., 1988.
- Benham, C., Mechanics and equilibria of supercoiled DNA, in Mathematical Methods for DNA Sequences, Waterman, M. S., Ed., CRC Press, Boca Raton, Fiz., 1988.
- Holley, R. W., Apgar, J., Evereti, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswich, J. R., and Zamer, A., Structure of a ribonucleic acid, Science, 147, 1462, 1965.
- Klm, S. H., Suddoth, F. L., Quigley, G. J., McPherson, A., Susanan, 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.
- Theses, I., Uhlenbech, O. C., and Levine, M. D., Estimation of secondary structure in ribonucleic acids. Nature (London), 230, 362, 1971.
- Stelm, P. R. and Waterman, M. S., On some new sequences generalizing the Catalan and Motzkin numbers, Discrete Math., 26, 261, 1978.
- Watermann, M. S., Secondary structure of single-stranded nucleic acids, Stud. Found. Combinatorics Adv. Math. Suppl. Stud., 1, 167, 1978.
- Watermann, M. S. and Smith, T. F., RNA secondary structure: a complete mathematical analysis. Math Biosci., 42, 237, 1978.
- Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J., Algorithms for loop matchings. SIAM J. Appl. Math., 35, 68, 1978.
- 11. Sankaff, 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., Melasner, F., and Hartmann, T., Compilation of tRNA sequences. Nucl. Acids Res., 132, 11, 1985.
- 15. Fox, G. E. and Worse, C. R., SS RNA secondary structure, Nature (London), 256, 505, 1975.
- Nishikawa, K. and Takemura, S., Structure and function of SS ribosomal ribonucleic acid from *Torulopsis* with J. Biochem., 76, 935, 1974.
- 17. Waterman, M. S., Computer analysis of nucleic acid sequences, Methods of Enzymology, in press.
- Trifonov, E. N. and Bolshol, G., Open and closed SS ribosomal RNA, the only two universal structures encoded in the nucleotide sequences. J. Mol. Biol., 169, 1, 1983.
- Woese, C. R., Magrum, L. J., Gupta, R., Slegel, R. B., Stahl, D. A., Kop, J., Crawford, N., Brosius, J., Gutell, R. R., Hogan, J. J., and Noller, H. F., Secundary structure model for hacterial 165 ribosomal RNA: phylogenetic, enzymatic and chemical evidence, Nucl. Acids. Res., 8, 2275, 1980
- 20. Noller, H. F. and Worse, C. R., Secondary structure of 165 ribusomal RNA, Science, 212, 403, 1981.

.

.

13

 Stlegler, P., Carbon, P., Zuker, M., Ehel, J. P., and Ehreamann, C., Structural organization of the 165 ribosomal RNA from E. coli iopography and secondary structure, Nucl. Acids Res., 9, 2153, 1981.
 Zwieb, C., Glotz, C., and Brimacoube, R., Secondary structure comparisons between small unit ribu-

- Zuieb, C., Clott, C., and aramacouse, n., accounty structure comparison comparison of the structure of the large somal RNA molecules from six different species, *Nucl. Acids Res.*, 9, 3621, 1981.
 Gluct, C., Zwieb, C., Brimacousbe, R., Edwards, K., and Konsel, H., Secondary structure of the large
- Glotz, C., Evree, C., Drumsremer, R., Eaveres, A., and Poster, H. Otoman and mouse mitochondrial subunit ribosomal RNA from Escherichia coli. Zoo mors chloroplass. and human and mouse mitochondrial ribosomes. Nucl. Acids Res., 9, 3287, 1981.
- Monzed, D., Stern, S., and Noller, M., Rapid chemical probing of confurmation in 16S ribosomal RNA and 30S ribosomal subunits using primer extension. J. Mol. Biol., 187, 399, 1986.
- Noller, H. F., Kop, J., Wheston, V., Brosles, J., Gutell, R. R., Kopylov, A. M., Dohme, F., Herr, W., Stabil, D. A., Gupta, R., and Worse, C., Secondary structure model for 23S ribosomal RNA. Nucl. Acids Res., 9, 6167, 1980.
- Kariku, S., Ost, F., and Blaladell, B. E., Patterns in DNA and aminu acid sequences and their statistical significance, in *Mathematical Methods for DNA Sequences*, Waterman, M. S., Ed., CRC Press, Boca Randon, Fia., 1988.
- 27. Waterman, M. S., Sequence alignments, in Mothematical Methods for DNA Sequences. Waterman, M. S., Ed., CRC Press, Boca Raton, Fla., 1988.
- Murlinez, H. M., An efficient method for finding repeats in molecular sequences, Nucl. Acids Res., 11, 4629, 1983.
- Galas, D., Eggert, M., and Waterman, M. S., Rigorous pattern recognition methods for DNA sequences. J. Mol. Biol., 186, 117, 1985.
- Badadur, R. R., Some Limit Theorems in Statistics. Society for Industrial and Applied Muthematics. Philadelphia, 1971.
- Ellia, R. S., Entropy, Large Deviations, and Statistical Mechanics, Springer-Verlag, New York, 1985.
 Pace, N., Olaen, G., and Woese, C., Ribosomal RNA phylogeny and primary lines of evolutionary descent, Cell, 45, 325, 1986.