Volume 12 Number 1 1984 Nucleic Acids Research

Generating non-overlapping displays of nucleic acid secondary structure

Bruce A.Shapiro, Jacob Maizel*, Lewis E.Lipkin, Kathleen Currey* and Carol Whitney*

Image Processing Section, Division of Cancer Biology and Diagnosis, National Cancer Institute, and *Molecular Structure Section, Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20205, USA

Received 15 August 1983

ABSTRACT

A new algorithm is presented which permits the display of nucleic acid secondary structure by computer. This algorithm circumvents the problem of overlapping portions of the molecule which is inherent in some other drawing programs. The results from this algorithm may also be used as input to the drawing algorithm previously reported in this journal [1] to untangle most of a drawing. The algorithm also represents the molecule in a form which makes visual comparisons for similarity quite easy since it guarantees that comparable features will reside in the same relative position in the drawings when the drawings are normalized.

INTRODUCTION

Programs and biochemical techniques [2-6] which permit the prediction and measurement of nucleic acid secondary structure have emphasized the need to visualize these molecules in a coherent manner. A few programs have been developed to accomplish this visualization [1,7-11] one of these was reported by us [1] (hereby referred to as the standard polygonal drawing). The published techniques share one drawback. Because of the geometric constraints that have been imposed by the drawing process, portions of the molecular drawing overlap. To alleviate this problem (which becomes more apparent as the size of the molecule increases) one must resort to an interactive process of removing the overlaps. This process is not unmanageable. Such an interactive solution has been reported in [1] where over 100 molecules have been untangled successfully. However, it would be more desirable if a method were available by which the secondary structure of a molecule could be depicted with minimal user interaction. This paper discusses such a technique and also shows how the technique may be applied to determining similarity of secondary

structure amongst several different molecules. The programs described in this paper accept region table input such as is described in [1]. A region table is a sequence of 4-tuples (start base position, stop base position, size and region stabilizing energy). The tuples must be presented in increasing order relative to the five prime position. The nested and branching structures which are inherent in secondary structure drawings are not directly apparent in the given region table. However, it is one of the purposes of the algorithms presented here to convert the given region table into graphical displays which make these structures quite evident. The programs permit the display of secondary stuctures on such devices as a Tektronix 4012, 4025, 4027 and VT-100 as well as producing hard copy on a plotter such as the Zeta.

CIRCLE GRAPH REPRESENTATION

As was discussed in [5,6] the secondary structure of a molecule may be depicted by a circle graph. An example of one such graph is shown in Figure 1. The primary sequence is layed out around the circumference of a circle. A chord is drawn joining every pair of bases that bond together as a result of a folding algorithm and/or biochemical data. It can then be seen that base paired regions of size n are represented by n parallel chords spaced one base apart. Other morphologic features can be seen in the circle graph. A hairpin loop is represented by the space between a chord and the circumference of the circle. A bulge loop is represented by the space between non-parallel chords one base apart on one side and m bases apart on the other, where m is the size of the bulge. An internal loop is represented by the space between two chords that may be parallel or non-parallel but are greater than one base apart on each end. A multibranch loop is represented by the space between at least three non non-parallel chords. In the circle graph representation it is assumed that the 5' side of the molecule starts at the top of the circle and that knots or crossing chords are not permitted.

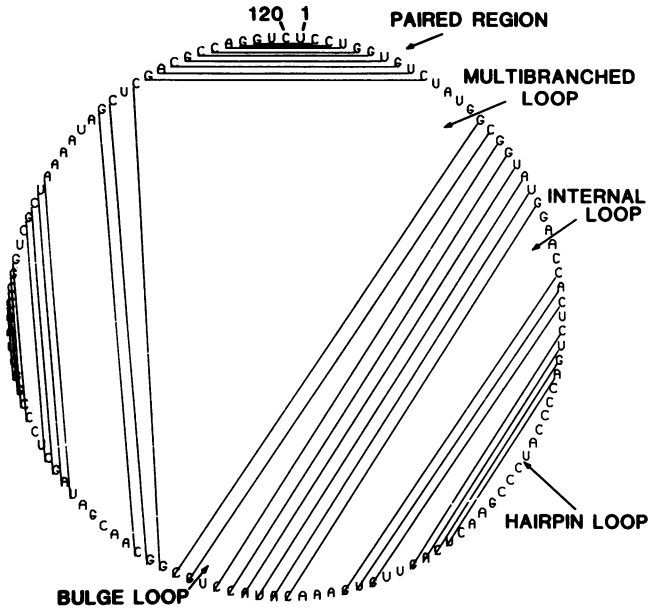

Figure 1. Circle graph of 5sRNA. Base pairing predicted by method of Zuker [6].

USING THE CIRCLE GRAPH TO GENERATE A SECONDARY STRUCTURE DRAWING

Given a circle graph as described in the previous section, it is possible to construct a secondary structure drawing (hereafter referred to as radial drawing) in the following way.

Everywhere a region exists (parallel chords one base apart) construct a radial line from the center of the circle bisecting the set of chords representing the region. The radial bisector will be perpendicular to this set of chords. The direction of this radial line defines the direction of a stem (REGION STEM) emanating from the circumference of the circle. Similarly, wherever a part of a loop appears (a contiguous set of bases along the circumference without any intervening chords) construct a radial line from the center of the circle bisecting the arc the ends of which are defined by the surrounding chords. The direction of the radial line defines the direction of another stem (LOOP STEM) emanating from the circle (see Figure 2a). The length of each stem is

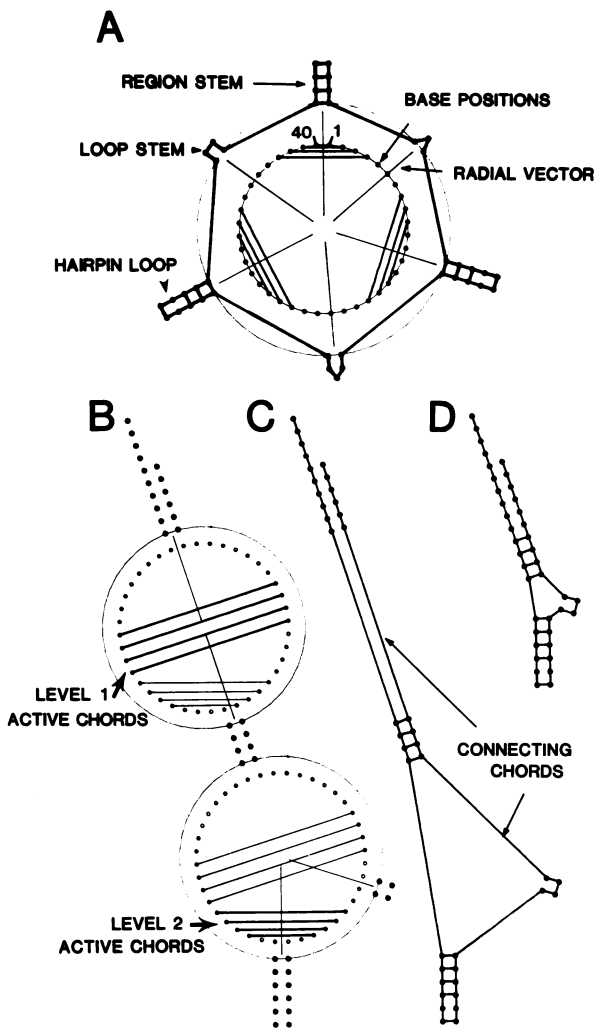

Figure 2. Construction of a radial drawing from a circle graph. A) A non-recursive structure; B) recursive structure showing two levels of circle and radial drawing; C) non-compressed radial drawing; D) compressed radial drawing.

determined by the number of bases that are found in that portion of the arc the stems represent. The LOOP STEM size will be proportional to the number of bases in the loop arc. A REGION STEM size will be proportional to the number of bases forming the region. In reality an internal loop will consist

of two stems, a multibranch loop will consist of n stems where n is the number of regions comprising the multibranch loop. Exceptions to this are those parts of a multibranch loop that have no intervening bases.

The simplest way of thinking about the positioning of the stems is to construct the circle with a circumference that is twice the size of the sequence. Thus, every two unit distances traveled on the circumference represents one base position. Depending upon whether the arc sizes under consideration are odd or even, the directional radius will lie on a base position or on an off base position (odd sizes will lie on a base position, even on an off base position). The width of a stem is defined to be two unit distances, one unit to each side of the direction radius. The length of a stem is the number of units equal to the number of bases in the arc under consideration. If the number of bases is even, the end of the stem will be square. If the number of bases are odd, the end of the stem will be triangular (see Figure 2a).

One may then arrive at a radial drawing by connecting chords (CONNECTING CHORD) to the ends of the stems at the points where they touch the circumference of the circle. They are connected in order of increasing base number with the last stem being connected to the first (see Figure 2a).

MULTIPLE LOOPS ON A STEM

In general, secondary structures are not as simple as those depicted in Figure 2a. Usually, there are chords within chords within chords, etc, as shown in Figure 2b. This degree of complexity is handled in a recursive fashion. Conceptually this is accomplished by constructing, for each level of recursion, another circle of the same radius as the first. The direction radius of the outer most set of region chords defines the position of the new circle, i.e. the angle is the angle of the direction radius extended to connect to the center of the new circle. The new circle is positioned a distance equivalent to the size of the stem (determined as described above, assuming that inner chords are non-existent) plus the radius of the circle. If another level is required beyond this, the same

process is continued as long as necessary, using the new outer most set of chords to define the new direction. The process continues for all sets of chords representing regions. Within each new circle, stems are generated as described for intervening loop arcs. The result of such a process is shown in Figure 2c. It should be noted that the ends of the stems are connected as described above.

COMPRESSING THE RADIAL DRAWING

The drawing may be compressed in size eliminating some of the wasted space created by the connecting chords that join the stems. This increases the resolution available for viewing these drawings. The shrinkage is accomplished by reducing the size of each circle at each level of recursion by the smallest distance between two adjacent stems on a given level. This has the effect of reducing the radius of each circle. The result of such an operation is visible in Figure 2d. It should be noted that the stem sizes increase proportionately with the decrease in radial size thus maintaining the same size stem amongst all the stems in the drawing.

DETERMINING BASE PAIRING FROM THE DRAWING

As can be seen from the illustrations, the technique used to form the radial image compresses some of the fine structure (loops, base paired regions). In order to visualize these structures and to determine base pairings a zoom facility has been incorporated into the program. If a particular part of the image is to be examined, a crosshair is positioned over the area (see Figure 3a). A zoom factor (e.g. 3X) is entered on the terminal. The resulting enlarged image indicates the base pairing by connecting lines (see Figure 3b). Where suitable hardware is available, a different color may also be used to display each of the four bases thus permitting rapid determination of base populations. After a zoomed subimage has been displayed the user than has the option of zooming another portion of the molecule by repeating this process. Figure 3c shows for comparative purposes the 5sRNA represented in the standard polygonal format.

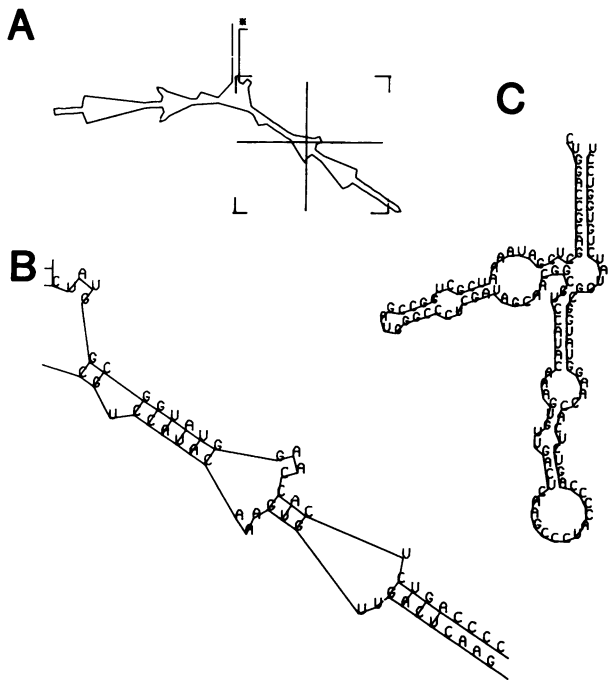

Figure 3. Comparison of radial and standard polygonal drawing of 5sRNA. A) Radial drawing (from circle graph of Figure 1a). An asterisk marks the lowest numbered base; B) zoomed (3X) radial drawing of boxed area of "a" above, illustrating bases and pairing; C) standard polygonal drawing.

AUTOMATIC UNTANGLING OF A STANDARD POLYGONAL DRAWING

As was reported in the previous paper by the authors one may generate a standard polygonal drawing which is more visually pleasing than the radial technique just described. However, as was reported it is necessary for a user to interact with the drawing to untangle the molecule. This process is not unmanagable but becomes more difficult as the size of the molecule increases. By using the angular information produced by the radial drawing technique described in this paper, one can almost completely untangle the standard polygonal representation automatically. It is then only necessary to interactively untangle the remaining tangled portion of the molecule and/or orient the molecule in such a way as to make the drawing more presentable for publication or similarity comparisons.

Since, the radial drawing technique guarantees non-overlaps, one may use the angles determined for the region stems in determining pivot angles for the stems in the standard polygonal drawing. As one may recall from [1], potential pivot points were defined to be at the base positions where a stem terminates. Previously, by pointing to a potential pivot point one was able to pivot by 0.5 radians each iteration. With a slight change to the algorithm one may read in pivot angles other than 0.5 radians for a base. This pivot angle is determined by the radial angle for a stem. The radial angle for a stem is defined as an absolute angle measured in a clockwise direction starting at the top of the circle graph representing the radial drawing. Every angle in the standard polygonal drawing is also converted to an absolute angle. The algorithm then determines for each pivot point the difference between the absolute angle that it is currently at and that which it should be at based upon the radial drawing. This difference angle then becomes the pivot angle for the base in question. This process is repeated for each potential pivot point (in our case only those bases that comprise the 5' most base of a stem and the 3' most base of a stem are used). Each time a pivot is made new absolute angles in the standard polygonal drawing are recomputed since a pivot in one base will alter the absolute angles of those bases that follow and that are bounded by the base opposite the pivoting base. At the end of the process an almost completely untangled molecule remains. Results of this process are shown in Figure 4. Figure 4a illustrates a mouse beta globin major precursor RNA in its original tangled form. Figure 4b depicts the same molecule after the automatic untangling process has been applied. Figure 4c shows the final presentation of the molecule after manually altering a few points.

It should be noted that for molecules greater than 500 bases in length the complexity of the structures generated make totally manual untangling somewhat difficult. However, we have found that automatically untangling the molecule first shortens the interaction time by a factor of about six.

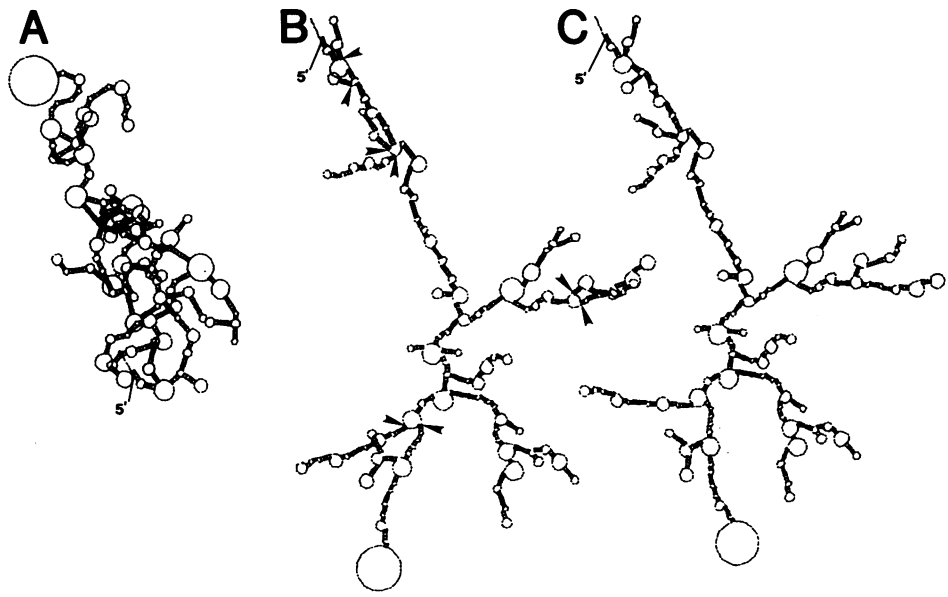

Figure 4. Automatic untangling of a standard polygonal drawing of mouse beta globin major precursor RNA [12]. The predicted structure of 1394 bases required seven hours of cpu time on a VAX 11/750 with 8 M bytes of memory using Zuker's FORTRAN program. A) Original drawing; B) automatically untangled drawing; C) final drawing after interative rotations at eight pivot points.

SIMILARITY DETERMINATION USING THE RADIAL REPRESENTATION

One of the powerful results of the radial representation is the use of this technique to visually determine similarity amongst a set of secondary structures. If one has two non-identical sequences of the same size and there are secondary structures within the total stucture that are identical or almost identical, then these structures will appear as identical or almost identical structures amongst the several drawings. The relative directions and positions of the structures will be preserved. The drawings may be rotated and scaled to facilitate comparisons. The algorithm currently permits the extension of the sequence at the 3' end with effectively "don't care" bases to permit the equalization of two non-equal length sequences. An addition is being considered to permit padding at arbitrary positions to accomodate splicing. Also, a rotation facility exists that

permits the user to specify the amount of rotation desired in the diagram by entering a base number or an angle. The base number defines the rotation angle by using the circle graph where the standard orientation (0 degrees rotation) is with the first 5' base at the top of the circle. The base number is converted into a rotation angle by defining the rotation amount as the angle between the radius of the circle graph drawn to the first base position and the radial line to the specified base position. The rotation facility is especially useful to compare two folded subsequences from the same sequence where the subsequences overlap each other. It is then possible to align the two segments by padding one subsequence, if necessary, to ensure equal lengths, and then rotate one subsequence with respect to the other to ensure exact alignment.

Once the desired rotation is accomplished, it is advantageous to scale the drawings to the same size viewing box. This facilitates comparisons since it makes the drawings more nearly the same size. Generally speaking, the viewing window is computed to be the smallest possible square box that will enclose the drawing. Thus, if two images have different size boxes, the scaling process increases the size of the smaller box to be the same size as the larger.

Since two structures are not necessarily identical when doing a comparison, the compression scaling, as described above, may have a tendency to compress features in one drawing and not in another because one drawing may not have a structure that the other one has. Thus, another version of the algorithm exists which does not make use of the compression scaling factor. This minimizes differences in the distance between stems when performing comparisons. This is shown in Figure 5.

Simliarity comparisons may also be made using the standard polygonal drawing technique which has been automatically untangled by the procedure described in this paper. Since molecules that have similar secondary structure will tend to have similar pivoting patterns, this approach appears quite viable. An example of this is shown in Figure 6.

Figures 6a and 6b depict predicted structures for mouse

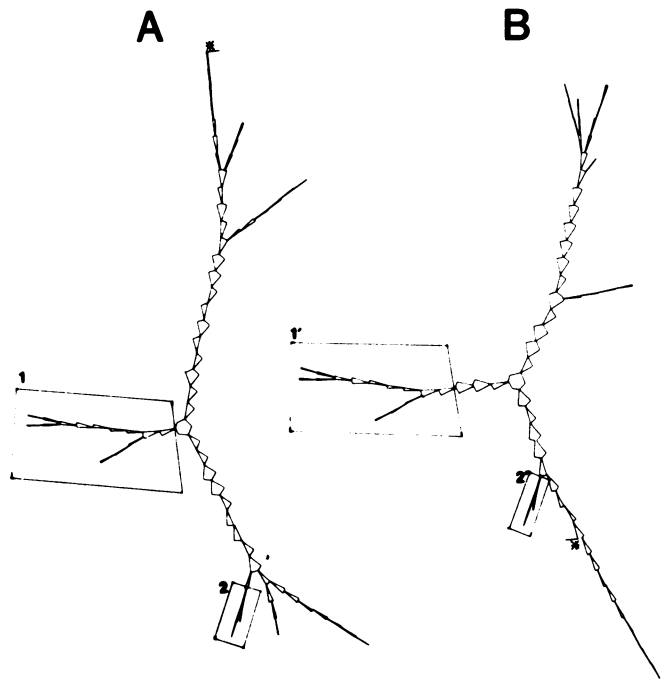

Figure 5. Use of non-compressed radial drawing to compare regions of similar hypothesized structure in overlapping segments of adenovirus 2 sequence [13]. Areas of identical substructure are enclosed in boxes (1-1', 2-2') after rotating drawing by 500 bases. A) Bases 3500 to 4500 were folded by the Zuker method; B) bases 4000 to 5000.

beta globin major precursor [12] and messenger RNA respectively. Although 709 of the 1395 nucleotides of the precursor RNA are spliced out to form the mature message, one notes a remarkable preservation of structure. Of the 184 base pairs present in the spliced message, 162 (88%) are conserved from the precursor RNA. Further examination of the drawings reveal that both intervening sequences (see Figure 6a) form nearly discreet secondary structures. Even though the splice junctions do not occur precisely at the base of stem structures, they are in close proximity in similar regions (in or near multibranch loops). This observation supports the possibility of secondary structure playing a role in splicing since the coding/intervening sequence (exon/intron) boundaries are so clearly defined. Among previous approaches to

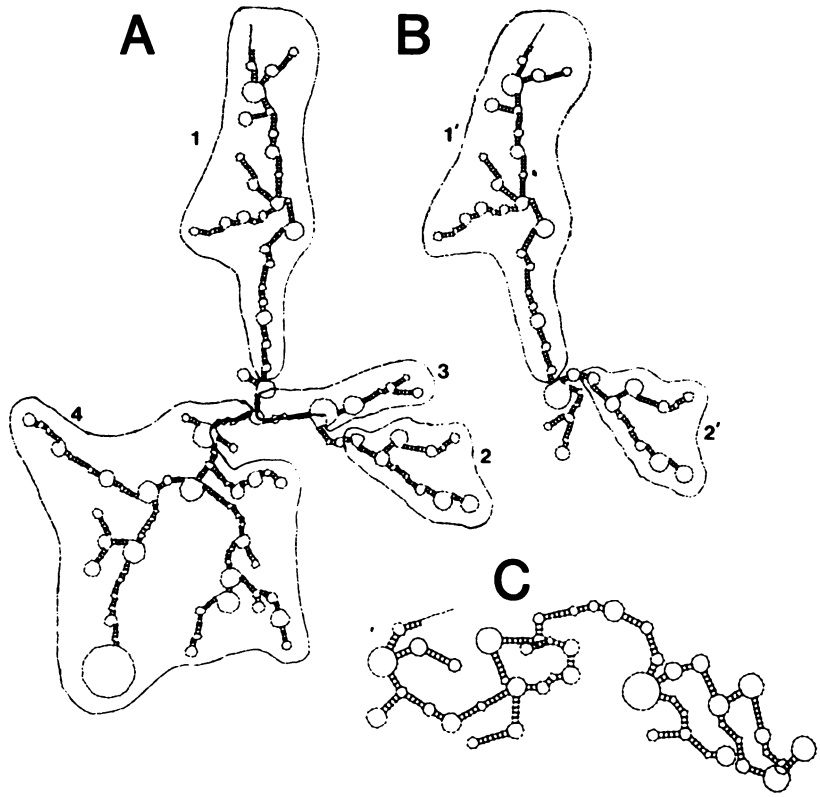

Figure 6. Use of the polygonal drawing algorithm to show similarity of predicted structures in precursor and message coding sequences. Areas 1-1' and 2-2' show comparable coding regions in precursor and message. Areas 3 and 4 show the clustered structures within intervening sequences 1 and 2 respectively in the precursor RNA. A) Precursor; B) message; C) message in original polygonal form. Compare this with Figure 4a.

this problem searches for dyad symmetry revealed very little, yet the first intervening sequence was observed as a discreet secondary structure in R-loop studies (Tilghman, et. al. [14] and J.V. Maizel, unpublished data). Go [15] has correlated DNA exonic regions with protein structural units in globin and Craik, et. al. [16,17] have shown that intron-exon splice junctions map at protein surfaces implying a "relation between the intron-exon structure of the gene and the tertiary structure of the gene product". The data of Figure 6 further the thought that yet another level of nucleic acid structure

may be correlated with domains at the protein level. Work is currently underway in our laboratory to substantiate, develop, and further refine this observation.

Figure 6c represents the beta globin message as drawn by the standard polygonal program without the input of the radial drawing algorithm. Even though this molecule could be easily untangled manually with only a few pivot points, it is apparent that similarity comparisons are more difficult between it (Figure 6c) and the completely untangled version (Figure 6a), and impossible between (Figure 6c) and the precursor drawn by the standard polygonal program (Figure 4a).

DISCUSSION

As can be seen from the preceeding discussion, the algorithm presented here permits the computer drawing of the nucleic acid secondary structure without the topological problems inherent in other algorithms which cause overlap. It has also been shown how the radial drawing algorithm can be used to almost completely untangle a drawing presented in the standard polgonal format. This method often produces results suitable for quick analysis even without manual pivoting. The radial algorithm has the added advantage of producing essentially a canonical representation of the molecule for the given structure. This permits various secondary structures to be compared visually. It should be pointed out that even though the standard polygonal drawings are more visually pleasing, i.e. allowing easier recognition of stems and loops and thereby incorporating a richer collection of visual cues, the radial drawings are more mathematically precise for similarity comparisons since the angle of the stems and loops are precisely defined. On the other hand, in the automatically untangled drawings using the standard polygonal format, the stems and loops may not be at precise locations due to modifications by the user and/or because of certain geometric and topological problems.

It should be mentioned, that the circle graph is probably the form that is most canonical and therefore most suitable for visual similarity comparisons. This

representation has been used by the authors to show changes in secondary structure as a molecule is synthesized from the 5' end by depicting a series of frames which were shown as a movie. The circle graph is still not as pleasing as the standard polygonal form nor does it depict features as well as either the standard polygonal form or the radial form. The radial form does require the use of the zoom facility to visualize small features.

A zoom feature has been added to the standard polygonal algorithm. This permits the user to interactively zoom in on an area to untangle. This greatly simplifies the interative portion of the algorithm especially when used in conjunction with the automatic untangler.

The algorithms described here are currently written in SAIL [18] and run on a DECSYSTEM-20 running the TOPS-10 operating system. A combined version of the standard polygonal algorithm and the radial algorithm has also been written in PASCAL [19].

REFERENCES

1. Shapiro, B.A., Lipkin, L.E., Maizel, J. (1982) Nucleic Acids Res. 10, 7041-7052.
2. Pipas, J.M. and McMahon, J.E. (1975) Proc. Nat. Acad. Sci. USA 72, 2017-2021.
3. Studnicka, G.M., Rahn, G.M., Cummings, I.W. and Salser, W.A. (1978) Nucleic Acids Res. 5, 3365-3387.
4. Waterman, M.S. and Smith, T.F. (1978) Mathematical Biosciences 42, 257-266.
5. Nussinov, R. and Jacobson, A.B. (1980) Proc. Nat. Acad. Sci. USA 77, 6309-6313.
6. Zuker M. and Stiegler (1981) Nucleic Acids Res. 9, 133-148.
7. Shapiro, B.A. and Lipkin L.E. (1983) in Computing in Biological Science (Eds. Geisow _Barrett), Elsevier Biomedical Press, 233-271.
8. Shapiro, B., Maizel, J. and Lipkin, B. (1981) Annals of the World Association for Medical Informatics, 4th Meeting, 93-98.
9. Auron, P.E., Rindone, W.P., Vary, C.P.H., Celentano, J.J. and Vournakis, J.N. (1982) Nucleic Acids Res. 10, 403-419.
10. Feldman, R. Personnal Communication, National Institutes of Health.
11. Lapalme, G., Cedergen, R.J. and Sankoff, D. (1982) Nucleic Acids Res. 10, 8351-8356.
12. Konkel, D.A., Maizel, J.V., and Leder, P. (1979) Cell 18, 865-873.
13. Roberts, R. Personnal Communication, Cole Spring Harbor.
14. Tilghman, S.M., Curtis, P.J., Tiemeier, D.C., Leder, P. and Weissmann, C. (1978) PNAS 75, 1309-1313.
15. Go, M. (1981) Nature 291, 90-92.
16. Craik, C.S., Sprang, S., Fletterick, R., and Rutter, W.J. (1982) Nature 99, 180-182.
17. Craik, C.S., Rutter, W.J., and Fletterick, R. (1983) Science 220, 1125-1129.
18. Reiser, J.F. (1976) SAIL User Manual, Stanford University Artificial Intelligence Laboratory memo AIM-289. Also available from U.S. Dept. Commerce. Nat. Tech. Info. Serv. No. AD-A045-102, Springfield, Va.
19. Jensen, K. and Wirth, N. (1978) PASCAL User Manual and Report, Springer-Verlag, New York.