Volume 10 Number 15 1982 Nucleic Acids Research

A high speed, high capacity homology matrix: zooming through SV40 and polyoma

James Pustell and Fotis C.Kafatos

Department of Cellular and Developmental Biology, The Biological Laboratories, Harvard University,
16 Divinity Ave., Cambridge, MA 02138, USA

Received 14 May 1982; Accepted 25 June 1982

ABSTRACT

We present a new homology matrix program which owes its basic conception to the two-dimensional dot matrices previously described (1,2), but has important improvements and new features. It scores sequence homology over an adjustable range and plots the scores which are above an operator-determined filtration level. Its powerful noise-filtration system, capacity for compression without much loss of information, and speed of execution make this program a valuable tool in the analysis of homologies, internal direct repeats and reverse repeats, including palindromic sequences. The properties of the program are exemplified by analysis of SV40 and polyoma DNA sequences.

INTRODUCTION

The recent, rapid accumulation of nucleic acid sequence information has necessitated the development of computer programs for analysis and visual display of that information, especially when comparing long sequences for homology. Two major approaches have been used. One, typified by the Korn-Queen (2) and Seq (4) programs, is to list all sequence segments which are thought to have significant similarities. The two sequences being compared are searched for matches as they successively "slide" along each other, assuming all possible alignments. For acceptance, matches are constrained as desired (e.g. according to number of bases which must be matched or percentage of mismatch tolerated, extent of gaps, etc.), and their listing is accompanied by estimates of statistical significance. Fundamentally, this approach is designed to focus on the trees rather than the forest. It has high precision and is ideally suited for locating specific subsequences (e.g. TATAA boxes), but it does not attempt to display the matches in the context of an overall comparison

0305-1048/82/1015-4765$2.00/0

between the sequences. Furthermore, it is cumbersome for analysis of internally repetitive sequences, and slow (expensive) when long sequences are involved. The second major approach is to display homology in two dimensions. One sequence is aligned on the X-axis of a sequence matrix, the other is aligned on the Y-axis, and dots are placed at X-Y coordinates corresponding to matches. In the simplest dot matrix (1), dots are placed wherever a base on the X-axis matches a base on the Y-axis. Rows of dots forming a 45° diagonal indicate a region of homology, lateral displacements of the diagonal correspond to deletions or insertions, and parallel lines suggest duplications. The enhanced version of the "graphic matrix" introduced by Maizel (2) and the "best fit" matrix of Hunkapiller (5) use various filtration methods to highlight the most significant matches (see below). The important feature of all matrix methods is that they display the regions of homology in context, i.e. that they reveal the distribution of matched and non-matched segments along the sequences. Precise identification of the matched sequences requires some effort, and the statistical significance of the matches is not immediately obvious, but the overall picture of the forest is communicated. Furthermore, matrix procedures are well-suited to display internal repeats, and they are relatively rapid and economical.

In our work we have found it helpful to first gain an overall appreciation of the similarities between two sequences by matrix plotting, and then progressively zoom in on segments that appear to be of special interest, ultimately analyzing them by a Korn-Queen or Seq-type program. We have developed programs which are appropriate for this approach, and which we expect to be of considerable general interest. In outline, our matrix program has the following features:

1) It filters noise by a weighted exponential curve across up to 100 adjacent bases -- greatly improving signal-to-noise ratio.

2) It uses letters as symbols, rather than dots, thus permitting discrimination between matches of different significance. Incidentally, this eliminates the need for a plotter: the output is created by a regular printer.

3) It permits compression of the matrix by as much as twenty-fold, with very little loss of information and essentially no increase in noise. As many as 3,000 bases can be analyzed across a page, and thus the overall features of comparisons between even very long sequences can be easily appreciated.

4) The degree of matrix compression, the extent of the sequence segment to be analyzed, and the noise suppression, are operator-set variables. Accordingly, it is extremely easy to begin with the overall analysis then to "zoom in" for examination of a shorter segment at higher resolution.

5) The matrix plot is accompanied by a grid system which greatly facilitates locating matched segments. At highest resolution, the sequences can be read directly from the x and y axes.

6) Appropriate settings permit identification of various types of homologies: direct repeats, inverted repeats, (regions of potential secondary structure), or palindromes.

7) Despite sophisticated noise filtering and information display, the program runs extremely fast and uses very little memory.

8) Ancillary programs permit listing of aligned segments. A consensus sequence can be produced from several aligned segments. These listings can be supplemented by estimates of significance, obtained from a Korn-Queen type program.

RESULTS AND DISCUSSION

Overall comparisons: SV40 and Polyoma

As an example of the capability of our program for overall comparisons, we present in Fig. 1 a comparison between the DNA sequences of polyoma and SV40 viruses (6). The entire 5,292 x 5,243 matrix is compressed twenty-fold so it fits into 1 and 3/4 pages. Filtering is set for a minimum homology score of 51%. These settings are discussed below. The diagonal segments clearly indicate recognizable homologies in the following regions (approximately identified relative to the numbering of the polyoma sequence): origin of replication (1-60 bp), region coding for the NH_2-terminal portion of all T antigens (180-360 bp), regions coding for the central (1400-2060 bp) and COOH-

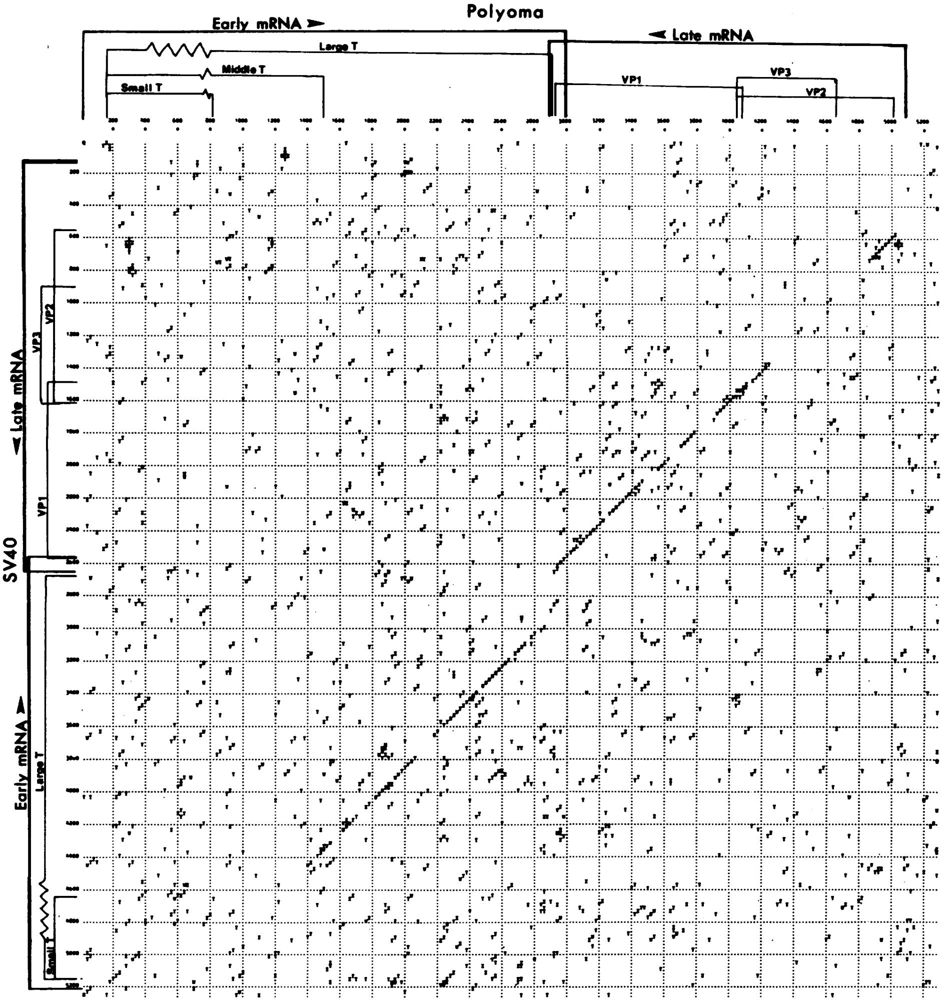


Figure 1 Matrix plotting of sequence homologies between Polyoma and SV40. A reverse matrix was used to retain the published numbering of the Polyoma sequence (6). Thus the diagonal differs by 90 degrees from the rest of the matrices shown and from the similar matrix of Maizel and Lenk(2). Thick lines show areas covered by the major RNA species. Thin lines indicate protein coding regions and zigzag lines represent their interruption by introns. Introns outside the coding region are not shown. Range=20 (41 bp window), Compression=20, Minimum Value Plotted=51%.

terminal (2180-2860 bp) portions of large T-antigens, region coding for the VPl protein and the COOH-terminal portion of the VP2 and VP3 proteins (2920-4240 bp), and region coding for the NH_2-terminal portion of the VP2 protein (4860-5020 bp). Comparable homology is not evident in the regions coding for much of middle-T antigen, which is not produced by SV40 (380-1380 bp), a short segment between the central and COOH-terminal portions of large-T antigen (2080-2160 bp), and much of the sequence of VP2 and VP3 proteins (4260-4840 bp). The shift in the diagonal at approximately 2930 bp corresponds to absence of a sequence which separates the coding regions of large-T antigen and VPl protein in SV40. These are indeed the prominent features identified by detailed comparison of the SV40 and polyoma sequences (7).

As a second example, we present a comparison of the SV40 sequence with itself (Fig. 2). In this case, the perfect diagonal row of A's is the line of identity, whereas structure in the pattern near that diagonal corresponds to internal repeats. Below we shall focus on the repeats between 0 and 300 bp which includes the region of 72 bp repeats (8).

Filtering

The most severe problem of matrix plots is noise. Since DNA has only four types of bases, simple dot matrices, based on single-base matches, will be at least 25% dots, even for unrelated sequences (higher if the base composition is skewed). Thus, all but the most striking homologies can easily be buried in the background noise (1,9).

A number of filtering methods have been used to screen out noise from dot matrices (2,5,10). Generally, they involve comparisons between small groups of bases rather than single bases: for example, a dot may be scored if 2-3 consecutive bases are identical (10), or if the percentage match of up to 10 consecutive bases exceed a certain minimum (2). The sequence alignment is then shifted by one base, and scoring is repeated. Noise is eliminated rapidly as the size of the group of bases being compared at a time increases; unfortunately, computation time also increases dramatically.

A more sophisticated filter was introduced by Hunkapiller in his "best-fit" matrix program (5). Here, matches are assessed

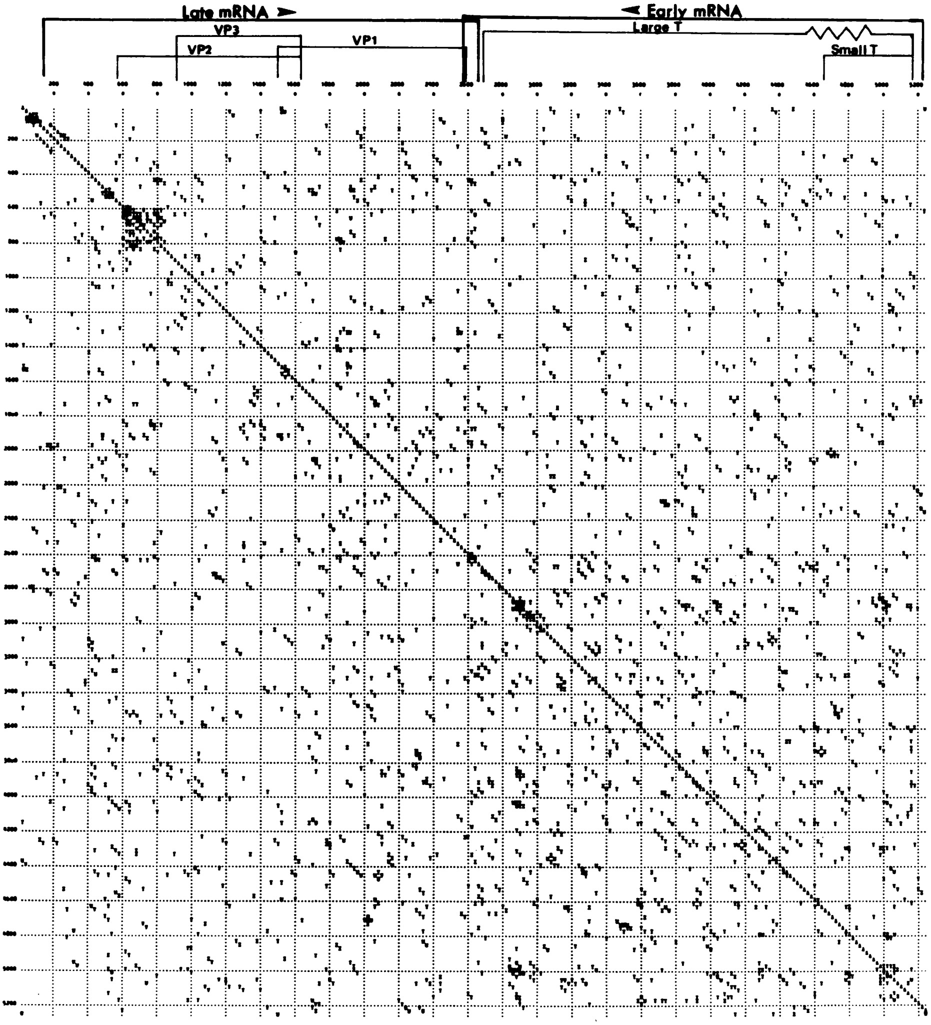


Figure 2 Self comparison of SV40 sequence, identifying internal repeats. Regions of genome as in figure 1. Range=20, Compression=20, Min Val=51%

over a long span centered on any two bases at a time. A homology score is computed, in which the contributions of progressively more distant matches are given relatively less weight in proportion to their distance from the aligned base pair. Thus, matches close to the aligned pair are most important, but the overall context also influences the estimate of fit; random local matches and mismatches are well suppressed. The "best-fit" alignment is displayed by placing only one dot per base, at the cordinate corresponding to the highest homology score for that base. Although time-consuming, this approach is well-suited to discovering distant or diffuse homologies.

A problem related to noise suppression is presentation of alternative significant alignments. This inevitably adds noise, but is essential if one wishes to detect internal repeats. The "best-fit" approach (5) discards all but the highest-scoring alignments, and thereby abdicates analysis of repeats. Other matrix programs display all alignments that are above a certain minimal cut-off. If a color plotter is available, matches of different significance level may be color-coded (2), but this approach has obvious limitations.

Our program achieves powerful noise filtration by exponentially weighing matches over long segments, as in the "best-fit" method. It displays all acceptable matches rather than the best fit, thus permitting analysis of internal repeats. By using letters as symbols rather than dots, it easily conveys an indication of the relative strengths of alternative matches. It permits adjustments in the shape of the homology scoring curve, and in the range of bases over which it operates; thus, parameters can be optimized for different tasks, from searching long stretches of sequences for homology to closely analyzing short internal repeats.

The program has been engineered to execute the comprehensive noise filtering extremely fast. Computers require at least four bytes to handle floating point numbers (fractions) and several operations to use these fractions, which are stored as a characteristic and a mantissa. Integers can require only two bytes and are handled by cpu registers in a straightforward manner. Our program inputs the operator set variables for the filtering curve, calculates all possible floating point values, converts them to integers, and stores the integer values <u>before</u> the matrix itself is calculated. Thus, the long loops which calculate the matrix values only perform comparisons and integer adding. Because of this optimization of program design to fit computer architecture, the matrix calculations are accelerated by as much as 50-70 fold. The program is further streamlined in a number of ways: for example, the forward and reverse matrix plots (useful for detection of direct and inverted repeats, respectively) are calculated by separate although almost identical programs. This saves two extra logical evaluations and therefore considerable time in comparisons involving long sequences. In the one case where execution time has been published (10), a standard dot matrix of 2 X 10^5 dots with a simple filter over 2-3 bases executes on a VMS/VAX with 7 minutes (real time) for computation and 13 minutes (real time) for writing the matrix on the disk. Our program, on the BBN C-70 (roughly equivalent to a PDP-11, a slower machine than a VAX), executes a larger 500 X 500 base matrix, using an exponential filter over 40 bases, in 9 minutes (real time) from start up to printed output. A noncompressed matrix can be computed as fast as it can drive a DEC LA-120 line printer. Because of this speed, the program can normally be run interactively, executing in real time and eliminating the need for large disk files or huge regions of dynamic memory.

For a given set of filtering parameters (shape and range of scoring curve), the program calculates the maximum possible score. Degrees of match are then expressed as percentages of this maximum score. These values are not strictly analogous to percent match, but are a good appoximation for the region close to the coordinate in question. A discrete value is calculated independently for every point in the matrix, and the value is encoded into a letter in 2% increments. Thus a 99-100% match is represented by "A", 97-98% by "B", 47-48% by "a", 45-46% by "b", etc. The operator also sets the minimum value to be displayed in the matrix at execution time, introducing another level of noise filtering. The program is fast enough so one can try a few values to find those which display the region of interest most clearly. A grid of co-ordinates is also printed in the matrix to provide a scale of reference. Thus, even with very complex sequence relationships, one can easily pick out matching regions, evaluate the relative strength of the match, and quickly locate them in the original sequence, from a single matrix. Use of letters as symbols permits the matrix to be displayed on a printer, eliminating the need for expensive digital plotters. Square matrices require a printer capable of short line feeds; with less capable printers the matrix emerges as a rectangle rather than as a square.

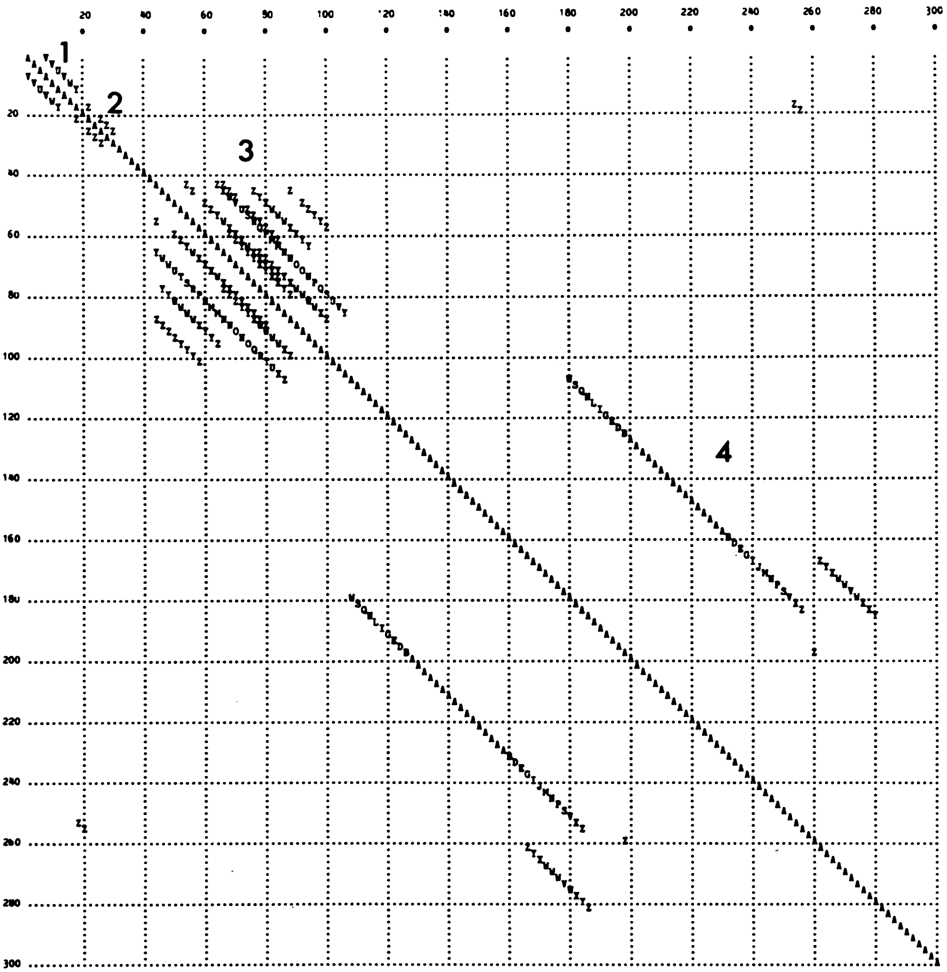
Compression and Expansion

When dealing with very long sequences, even streamlined programs are time-consuming. Furthermore, if the output is too large, visual analysis is hindered. Dot matrix programs may address this problem by assigning coordinates to groups of bases, rather than to individual base pairs (2). However, if evaluation is performed by averaging each group, information is suppressed. We have developed a procedure for compressing matrices with very little information loss and little increase in noise. As Figs. 1 and 2 demonstrate, compressed 5,000 x 5,000 matrices can be projected in a printed page and convey considerable information, although this degree of photographic reduction does make the letters difficult to read.

When we compress a matrix, we still calculate homology scores for individual base-pairs, rather than for groups. The scores of the bases to be grouped within a single coordinate are then scanned by the computer, and the highest value (not the average) selected to represent the group. Thus, the existence of an acceptable match is not masked by compression. An additional refinement that preserves information is the use of parallelograms rather than squares to group bases for compression. Since features of interest in a matrix are 45° diagonals, parallelograms avoid the artefacts that are introduced by grouping in squares, if a diagonal crosses through more than one square in a line or column.

When we compare long sequences for the first time, we use a high compression value and a relatively long degree of noise filtration, as in Figures 1 and 2. The main features are thus perceived, and we are in a position to "zoom in" on areas of special interest by expanding the matrix, and to clarify the matches by adjusting the noise filtration. Examples are shown in Figures 3 and 4.

Fig. 3 shows a expanded self-comparison of the first 300 bases of SV40 DNA, from the origin towards the late region; it

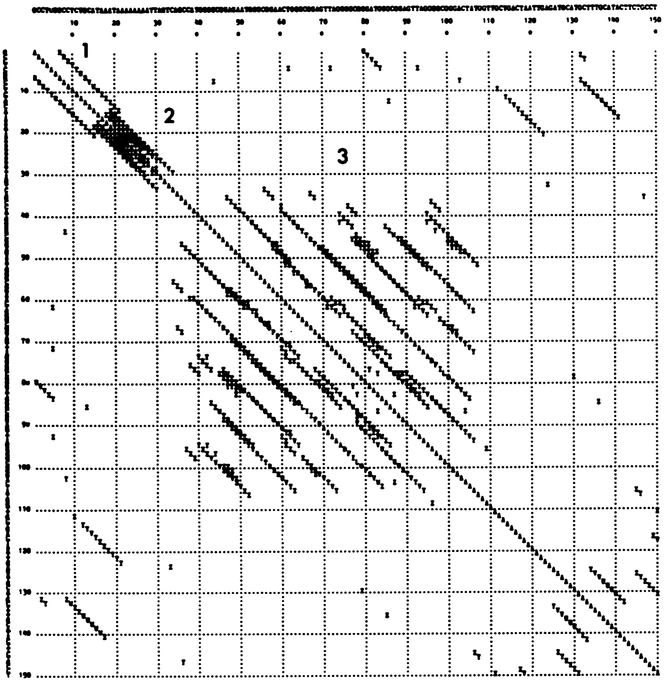


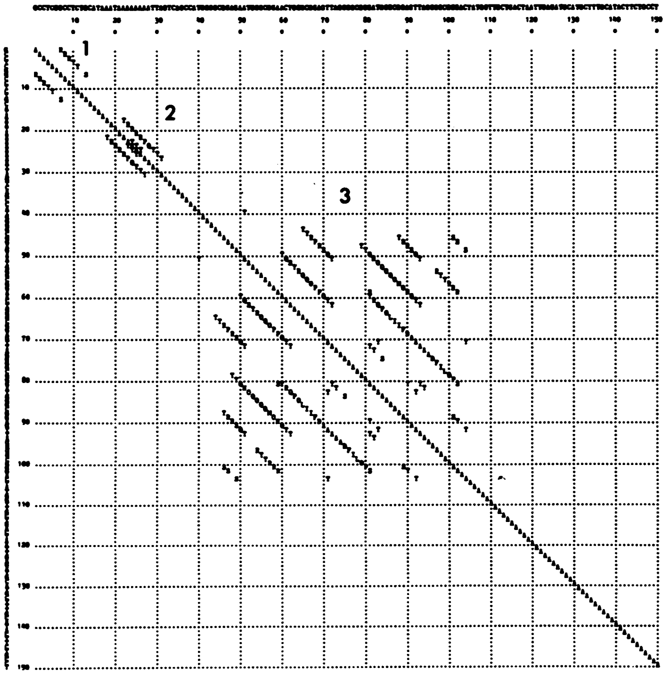
```
TGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTCCACACC
000000000000006000606000000000000000000000000000000000000000000600000006C
107                                                                      183
TGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTCCACACCTGGTT
TGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTCCACACC250
179       CTAAcTGAcAc                                             ACaTTCCACAgCTGGTT
          251    261                                              262            278
```

Figure 3 (Upper) Expanded region from the upper left corner of Figure 2; bases 1-300. Four regions of direct repeats are identified as features 1-4. Range = 20, Compression = 20 Min Val = 50%. (Lower) Output from ancillary sequence alignment program for feature 4. The first line is the consensus sequence (in this case set so more than 50% of the bases at any position must match it). Second line shows % of bases in aligned sequences which match the consensus at that position: 0 = 100%, 9 = 90%, 8 = 80%, etc. No number in the second line and a period in the first line would indicate no consensus is possible at that position for the % match required (cf. Figure 4). In the aligned sequences (third to fifth line), matches to consensus are in upper case and underlined, while mismatches are in lower case. The third and fourth lines correspond to the two long diagonals in feature 4; these are the 72 bp repeats (enhancer sequences). In the last line, the 262-278 bp segment corresponds to the short lines in feature 4 of the matrix; since it overlaps the end and the beginning of the consensus sequence, five bases (179-183; arrowhead) have been duplicated in one of the main repeats for optimal display. The other short segment of the last line, from 251 to 261 bp is only represented in this matrix by the single Z at co-ordinates 260;200. To see this match clearly, a shorter range must be used (not shown).

corresponds to the upper left corner of Fig. 2 (one and a half squares across). A prominent, well matched segment (indicated by 4), parallel to the line of identity, corresponds to the well known tandem 72 bp repeats (enhancer sequence). The repeat length can be easily measured along the X- or Y-axes, using the grid. A second, shorter line, parallel to the 72 bp repeat and displaced 22 bp relative to it indicates that the end of the 72 bp repeat is further reduplicated. Changing the parameters to highlight short matches (not shown) reveals duplication of another fragment of the 72 base repeat. The complete and partial repeats can be immediately seen by using an ancillary program, which lists and aligns the matched sequences (Fig. 3, lower). Additional divergent repeat fragments also exist beyond the region shown, in the 3' direction (data not shown).

Fig. 3 also shows the pattern of direct repeats (labelled 1, 2 and 3) closer to the origin. These repeats are examined at higher resolution in Fig. 4, which includes two non-compressed matrices, one at low filtration (Fig.4, top) and one at moderate filtration (Fig. 4, bottom). Non-compressed matrices are always accompanied by the sequence itself, and thus the nature of the repeats can be easily understood by reading and aligning manually the matched segments. Alternatively, the ancillary listing program can be used. If an estimate of significance of the match is desired, or if there is reason to believe that the optimal alignment includes gaps, the Korn-Queen program can be used at





```
                  GCCTC.
                  00000

  1      6  83.3% GCCTCg
  7     12  83.3% GCCTCt
                    1
```

```
                  ATGGGCGGAG.T
                  7500880856 0

 40     50  72.7% tgGGGCGGAGa
 51     60  90.0% ATGGGCGGAa
 61     72  83.3% cTGGGCGGAGtT
 73     81  77.8% AgGGGCGGg
 82     93  91.7% ATGGGCGGAGtT
 94    105  66.7% AgGGGCGGgacT
106    115  60.0% ATGGttGctG
                       3
```

this point.

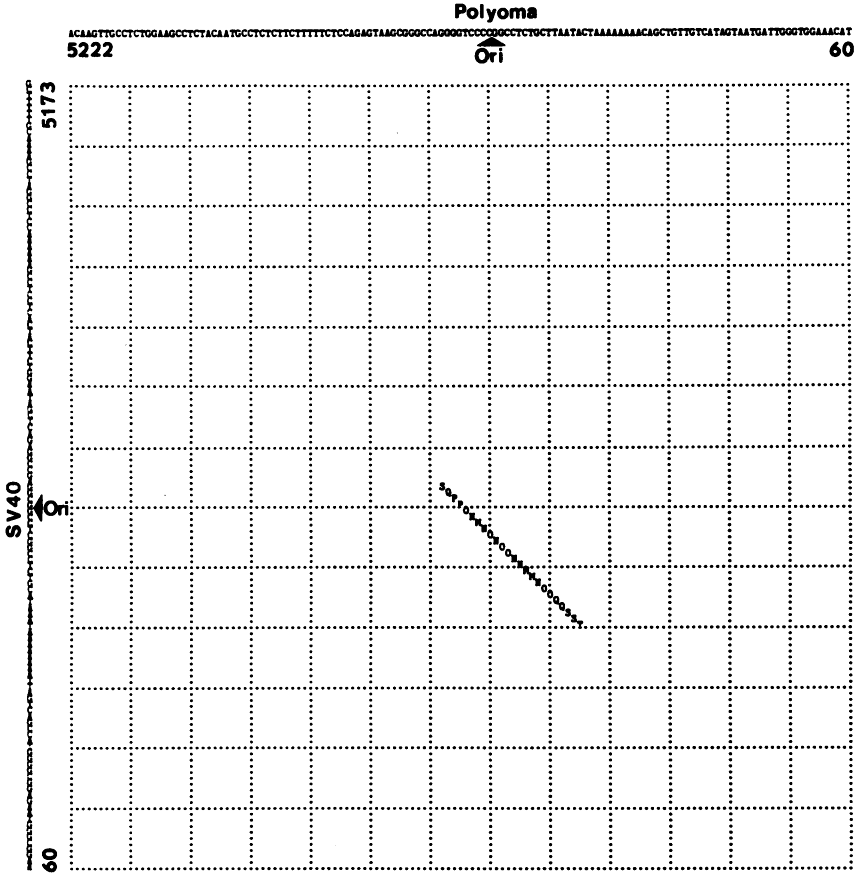
As Fig. 4 shows, feature 1 is caused by a short, tandem, almost perfect hexanucleotide repeat that begins at the origin. It is detected even in matrices which use a long range, because the SV40 sequence is circular, and the program can "go around the end." With linear sequences, loss of information near the ends is unavoidable, at distances that depend on the range used for scoring. Feature 2 corresponds to the A-rich region, which includes the Hogness box for early transcription (towards the origin), and appears as a repeat because of matching of staggered oligo A stretches. These two features are too limited to be seen in the 20-fold compressed matrix of Fig. 2, but they do appear in two-fold and five-fold compressed matrices (Fig. 3 and data not shown).

Feature 3 can be seen at low noise suppression to consist of six repeats, spaced at an average period of 11 bp (Fig. 4, top). The diagonals often consist of two or three lines, offset by one or two boxes. This is an indication that the repeats include gaps. The ancillery listing program reveals (Fig. 4), that the repeats include the invariant sequence GGGCGG, and differ in length (10 to 12 bp); a seventh, less well matched repeat is also detected. At higher filtration (Fig. 4, bottom), a strong 21 bp periodicity becomes apparent, which is due to the near identity of the third repeat with the fifth, and of the fourth repeat with the sixth.

By a similar process of examining progressively more expanded and noise-suppressed matrix plots, we have considered the homologies between polyoma and SV40 DNA in the 450 bp that surround the origin of replication. By far the most significant

Figure 4 Region from the upper left of Figure 3, bases 1-150. Range = 8, Compression = 1. Note the effect of filtration on demonstration of direct repeats (features 1,2 and 3). (Upper): With less noise suppression (Min Val = 50%) feature 3 is seen to consist of six aproximately 11 bp tandem repeats. (Lower): At a higher supression (Min Val = 62%) optimized alignments of the repeated segments can be read from the matrix. (Right): Alignments of repeats in features 1 and 3, as generated by the ancillary alignment program. The consensus and percent match (see Fig. 3) are shown above the aligned sequences and, on the left, the first and last bases are numbered and the percent match listed for that line.

match includes the SV40 origin itself. This analysis (Fig. 5) suggests that the conventional origins (defined by convenient restriction sites in the two viruses) are displaced relative to each other by 5 bp.



```
...T..TT.T........T.AG.GG.C.AGG.G.CC.CGGCCTCTGC.TAA....AAAAAA.....C.G...T.....
   0  00 0        0 00 00 0 000 0 00 0000000000 000    000000     0 0   0

cacTacTTcTggaatagcTcAGaGGcCgAGGcGgCCtCGGCCTCTGCaTAAataaAAAAAAttagtCaGccaTggggc
tctTctTTtTctccagagTaAGcGGgCcAGGgGtCCcCGGCCTCTGCtTAAtactAAAAAAaacagCtGttgTcatag
```

```
                 T.AG.GG.C.AGG.G.CC.CGGCCTCTGC.TAA....AAAAAA
                 0 00 00 0 000 0 00 0000000000 000    000000
                                Ori
SV40      5229 TcAGaGGcCgAGGcGgCCtCGGCCTCTGCaTAAataaAAAAAA 28
Polyoma   5273 TaAGcGGgCcAGGgGtCCcCGGCCTCTGCtTAAtactAAAAAA 23
                                 Ori
```

Reverse repeats and dyad symmetries

Reverse repeats, including dyad symmetries (non-alphabetical "palindromes") are of special interest in nucleic acid analysis: they may mark the sites of intra-chain base pairing in DNA cruciform structures or RNA secondary structure, or the sites of recognition by regulatory proteins. Recognition of reverse repeats is performed in an exactly analogous manner. Usually we prefer to use a dedicated program, which displays reverse repeats as segments perpendicular to the diagonals. As shown in Fig. 6, the sequence to be analyzed is entered on one axis as is and on the other axis as the complement (not the reverse complement), and the comparison is made in reverse. This ensures that the normal numbering will be maintained. Alternatively, a normal matrix comparison can be run between a sequence and its reverse complement.

Fig. 6 shows the prominent reverse repeats and palindromes characteristic of the SV40 origin. Palindromes are recognized as segments which cross the diagonal, whereas reverse repeats that might give rise to stem loops are seen as symmetrical segments at a distance from the diagonal (features 5-7 and 8, respectively). Again, the matrix plotting highlights these features and permits their early identification in the sequence. The most prominent palindrome (feature 5) pairs the previously noted hexapeptide repeats (feature 1, Fig. 4) with a complementary sequence, symmetrically located across the origin on the early side; it is followed by an A-rich sequence on the late side. A related but less perfect palindrome, similarly associated with an A-rich sequence, is also present in polyoma DNA and is the basis of the homology evident in Fig. 5. In SV40, two other features evident in a reverse repeat and dyad symmetry plot (features 6 and 8) involve the same circa-origin sequence in alternative mutually exclusive pairings (Fig. 6).

Figure 5 (Upper): Homology matrix plot of region around origin of replication (Ori) of Polyoma, compared to the same region of SV40. The regions compared are delimited by the base numbers at each edge of the matrix. Range = 15, Compression = 1, Min Val = 63%. (Middle) Alignment of the general region of match. (Lower): The parts of the aligned sequences we would consider significantly matched.

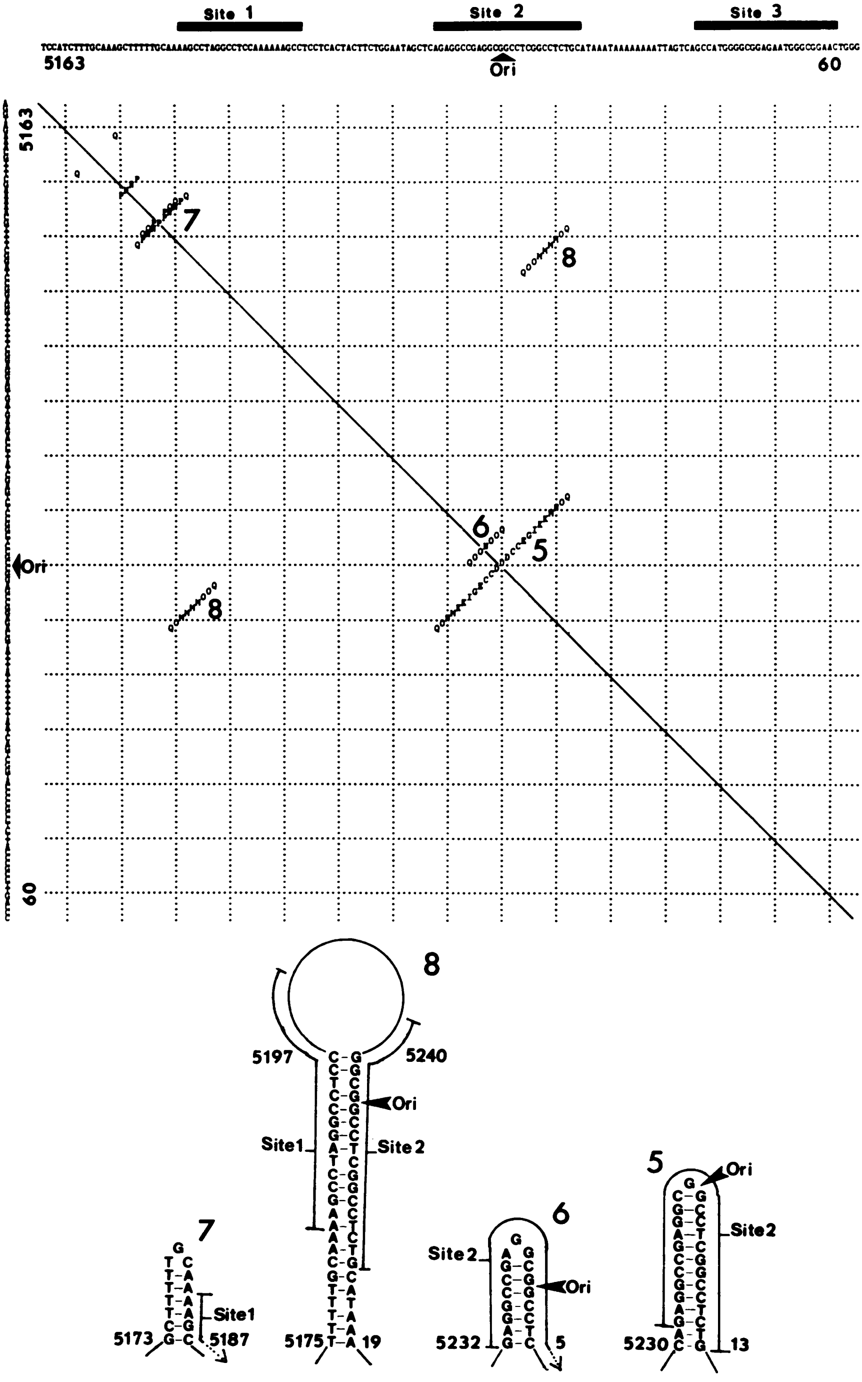

Site 1
Site 2
Site 3
TCCATCTTTGCAAAGCTTTTTGCAAAAGCCTAGGCCTCCAAAAAAGCCTCCTCACTACTTCTGGAATAGCTCAGAGGCCGAGGCGGCCTCGGCCTCTGCATAAATAAAAAAAATTAGTCAGCCATGGGGCGGAGAATGGGCGGAACTGGG
5163
Ori
60
5163
Ori
60
7
8
6
5
8
8
5197
5240
Ori
Site1
Site2
5175
19
7
5173
5187
Site1
6
Site2
Ori
5232
5
5
Ori
Site2
5230
13

Figure 6 (Upper): Reverse repeats and symmetries revealed by a reversed matrix plot of the region around SV40 origin of replication (Ori) on the X-axis, compared to its complement on the Y-axis. Range = 10, Compression = 1, Min Val = 68. Bars at top delimit T-antigen binding sites (11). Diagonal line is the axis of symmetry. (Below); Alignment of the self-complementary features revealed by the matrix.

It is intriging to note that the positions of T-antigen binding (11) are intimately associated with the potential secondary structures revealed by Figure 6. Palindromes 5 and 6 involve T-antigen binding site 2. Feature 8 is a stem loop involving both sites 1 and 2 almost symmetrically, and excluding loops 5 and 6. This series of alternate stem loops invites speculation regarding the molecular mechanisms of binding of T-antigen first to site 1, and then to sites 2 and 3 (12).

Program availability

The program is in Fortran 77 and can be easily adapted to any Fortran system. It is compatible with the DNA sequence package previously described (13). It is available as hard copy, on 8" single sided, single density, IBM format floppy disks, on tape, or over the modem at 300 or (preferably) 1200 BAUD. For obtaining the program, it is very helpful to send us specific information regarding your machine or system, e.g. the tape recording parameters, operating system, and billing information for the small charge levied for materials, mailing and handling.

Acknowledgements

This work was supported by NSF and NIH grants to F. C. Kafatos. We thank B. Klumpar for expert photography.

REFERENCES

1. Konkel, D.A., Maizel, J.V., and Leder, P. (1979) Cell 18: 865-873.
2. Maizel, J.V. and Lenk, R.P. (1981) PNAS 78: 7665-7669.
3. Queen, C.L. and Korn, L.J. (1980) Methods Enzymol. 65: 595.
4. The "SEQ" program, SUMEX System, Stanford University
5. Steinmetz, M., Frelinger, J.G., Fisher, D., Hunkapillar, T., Pereira, D., Weissman, S.M., Vehara, H., Natenson, S., and Hood, L. (1981) Cell 24: 125-134.
6. Appendices A and B *in* DNA Tumor Viruses (1980), John Tooze

(ed), pp. 799-896. Cold Spring Harbor Laboratory Monographs.

7. Soeda, E., Arrand, J.R., Smolar, N., Walsh, J.E. and Griffin, B.E. (1980) Nature 283: 445-453.
8. Banerji, J., Rusconi, S., and Schaffner, W. (1981) Cell 27: 299-308.
9. Heiter, P.A., Max, E.E., Seidman, J.G., Maizel, J.V., and Leder, P. (1980) Cell 22: 197-207.
10. Novotny, J. (1982) Nuc. Acids Res. 10: 127-131.
11. Tjian, R. (1979) Cold Spring Harbor Sym. Quant Biol. 63: 655-662.
12. Tjian, R. (1978) Cell. 13: 165-179.
13. Pustell, J. and Kafatos, F. (1982) Nuc. Acids Res. 10: 57-59.