© 1991 Oxford University Press Nucleic Acids Research, Vol. 19, No. 14 3907–3911

A sequence assembly and editing program for efficient management of large projects

Simon Dear and Rodger Staden*
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received April 26, 1991; Accepted June 19, 1991

ABSTRACT

We describe a sequence assembly and editing program for managing large and small projects. It is being used to sequence complete cosmids and has substantially reduced the time taken to process the data. In addition to handling conventionally derived sequences it can use data obtained from Applied Biosystems,Inc. 373A and Pharmacia A.L.F. fluorescent sequencing machines. Readings are assembled automatically. All editing is performed using a mouse operated contig editor that displays aligned sequences and their traces together on the screen. The editor, which can be used on single contigs or for joining contigs, permits rapid movement along the aligned sequences. Insertions, deletions and replacements can be made in individual aligned readings and global changes can be made by editing the consensus. All changes are recorded. A click on a mouse button will display the traces covering the current cursor position, hence allowing quick resolution of problems. Another function automatically moves the cursor to the next unresolved character. The editor also provides facilities for annotating the sequences. Typical annotations include flagging the positions of primers used for walking, or for marking sites, such as compressions, that have caused problems during sequencing. Graphical displays aid the assessment of progress.

INTRODUCTION

There is increasing interest in methods to aid large scale sequencing projects. During such work a significant fraction of time is spent using computers for assembling, checking and editing the data. Assembly need take almost none of the users time as it is readily automated, but checking and editing can still require a great deal of time and concentration. We report a program that can perform assembly efficiently and which contains a new editor for aligned sequences that greatly reduces the time and effort required by users during sequencing projects.

The program (xdap) is based on a previous program sap(1,2) which has already been used for many large sequencing projects including Epstein-Barr virus (172kb)(3), human cytomegalovirus (229kb)(4), and also one of 150kb done as a single shotgun (A. Davison, personal communication). Although there have been important improvements in the assembly algorithms, we concentrate here on describing the new contig editor and some new graphical displays.

An important new development in sequencing technology is the availability of automated gel readers, such as the Applied Biosystems 373A and the Pharmacia A.L.F. One major potential advantage of using such computer controlled instruments to determine sequences is that the primary data is stored in machine readable form and so can be examined by the user during the editing process. The new editor takes advantage of the availability of the primary data by enabling the user to display the traces, in colour, below the aligned sequences. This alone gives a great saving in time, especially as the traces are exactly aligned about the current cursor position. We emphasize, however, that the new editor does not depend on the use of data from sequencing machines. All its editing functions work using conventionally determined data and the program is also useful for processing such sequences.

The editor also contains functions to enable users to annotate features within the individual sequences. Such a facility is useful for marking primer sites used during walking phases of a project, or for indicating the positions of known problems within particular readings. Each type of annotated site has an associated highlighting colour which is used in the editor display to show the feature positions.

MATERIALS AND METHODS

Data collection and preprocessing

For illustration we describe the way in which the data is being collected and preprocessed during the C. elegans genomic sequencing project (5), but the program is fully compatible with alternative strategies.

The bulk of the sequencing in this project is being performed on ABI 373A and Pharmacia A.L.F. fluorescent sequencing machines, and the data transferred respectively from their Apple Macintosh and Compaq personal computers to Sun SPARCstations for processing. The whole of the machine data (not just the sequence) is transferred, unedited.

Data from the start of a machine run contains the primer site, and data collected near the end of the run is generally too poor for inclusion in the assembled sequence. Initial processing on the Sun includes the use of a trace editing program, ted(6) which

* To whom correspondence should be addressed

allows users to visually select left and right cutoff positions to denote the start and end of good data. Users may also edit the sequences at this point. The output from ted is a file containing the sequence and a header giving the name of the original primary data file plus the cutoff positions. ted is used in a script (R. Durbin, C Lee, unpublished) that also produces a file of file names for all the sequences processed in any particular run. From this point on all processing is performed by xdap: it assembles batches of readings, or aligns individual sequences, and it also contains all the editing and management functions to correct and manipulate the data.

The assembly process

To assemble a batch of sequences the user need only type the name of their file of file names, and supply a few parameters that control the matching and alignment process. Assembly is rapid, no user intervention is required, and any number of sequences can be processed in a single batch. Each new reading is compared in both orientations with a consensus of all the previous data and optimal alignments produced for any matching regions. Up to the best 10 alignments are sorted into score order.

For each reading five outcomes are possible. There will either be no overlap (i), a single overlap (ii), an overlap with two contigs (iii), more than one overlap with a single contig (iv), or more than two overlaps (v). In case (i) a new contig is started. In case (ii) the sequence will be added to the matching contig. In case (iii) the sequence will be added to one of the contigs, the consensus for that region recalculated and recompared with the other contig; if the match is still good enough the two contigs will be joined. Generally cases (iv) and (v) are rare and are caused by repetitive sequences. To deal with these cases the program will always add the new sequence only at the position it matches best. For all cases the new sequence is aligned with the contig by inserting padding symbols in appropriate positions. If necessary, contigs are complemented before being joined. The names of any sequences that give problems, such as cases (iv) and (v), are written to an error file. All alignments are displayed during processing.

During a large project, particularly in the early stages and when using a strategy including a shotgun phase, the user will repeatedly go through the steps just described. Occasionally xdap will report a possible overlap for which it could not produce a sufficiently good alignment, in which case some interactive editing would be required. Towards the end of a project, or during a more directed phase, far more use is made of the editor and the graphical displays.

Graphical displays of contigs

As an aid to assessing progress and to help plan experiments we have added functions that produce graphical representations of the contigs. They are best understood by referring to the examples shown in figure 1. The top level function draws a single line to represent each individual contig. The lines are drawn alternately at one of two heights: the first at height one, the second at height two, the third at height one, etc. The length of each line is proportional to the length of the corresponding contig.

At the next level is a function that produces a schematic of a single contig by drawing a horizontal line to represent each of its gel readings. The lines show the relative positions and lengths of each reading. The plot is divided into two sections by a horizontal line that is identified by an asterisk drawn at each end. All lines that lie above this line represent readings that are in their original orientation, all lines below show readings that are in the complementary orientation to their original.

At the bottom level is a function to display the 'quality' of the data for each point in a contig. The quality of the data depends on the number of times it has been sequenced on each strand and the particular uncertainty codes used in each gel reading(1). The data is divided into five categories each of which is given a pair of y coordinates: well determined on both strands and they agree (0,0), well determined on the plus strand only (0, −1), well determined on the minus strand only (0,1), not well determined on either strand (−1,1), well determined on both strands but they disagree (−2,2).

When the quality analysis is displayed graphically the following scheme is used. The x axis represents the length of the contig and the y coordinates take the values −2, −1, 0, 1 or 2. Using the coordinate pairs given above for the bottom and top y values, the quality codes attributed to each base position are plotted as rectangles. Each rectangle represents a region in which the quality codes are identical, so a single base having a different code from its immediate neighbours will appear as a narrow rectangle, but a whole region well determined on the plus strand only will appear as a rectangle with y values of 0 and −1. A rectangle with y values of (0,0) obviously appears as a single line along the midpoint, and represents a section of sequence that is well determined on both strands.

A crosshair function is available for use with the graphical displays. This allows the user to employ the mouse to position the crosshair above any of the displays and ask, for example,

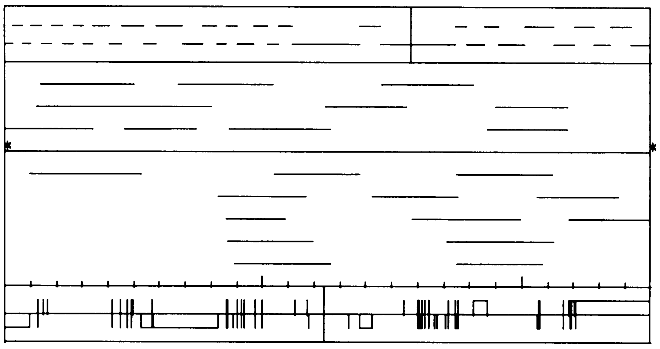


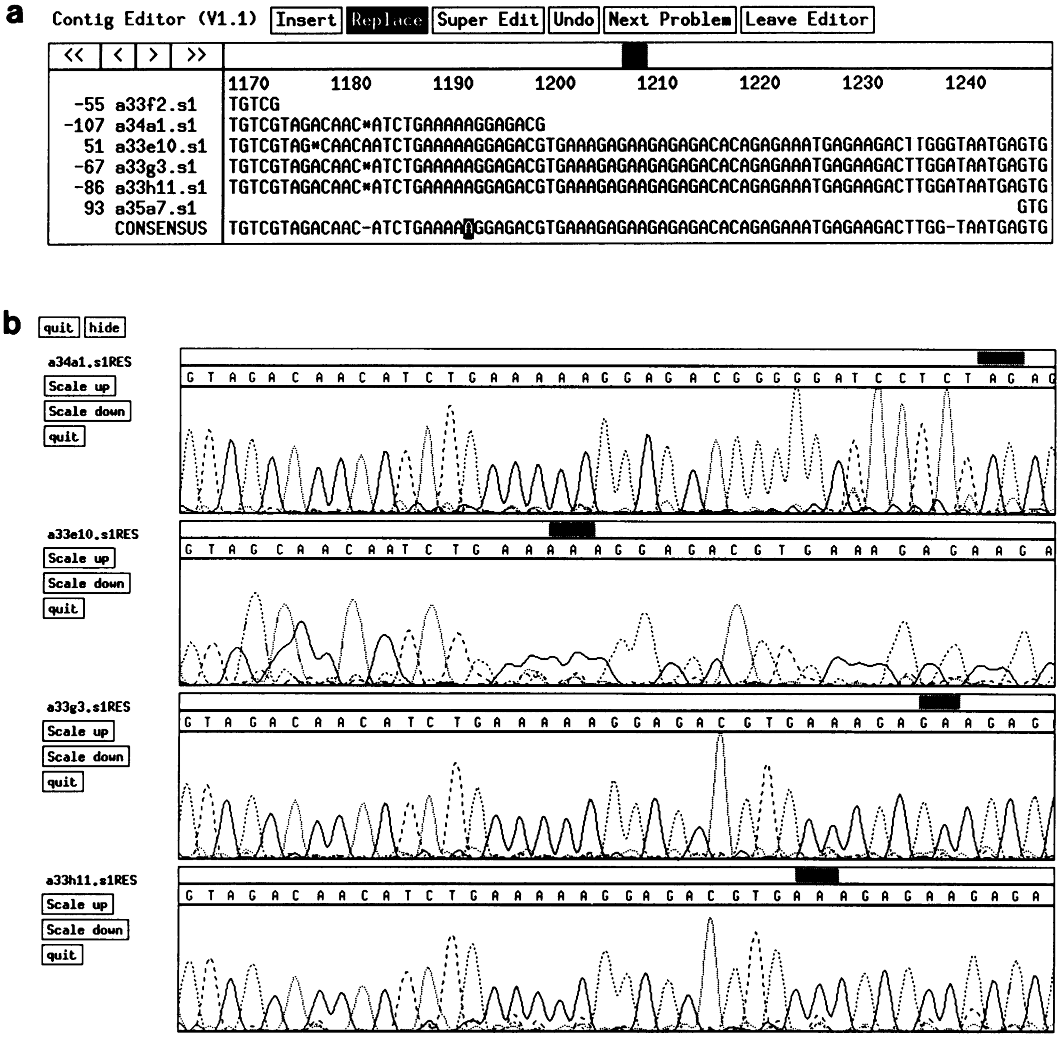
Figure 1 shows a typical graphical schematic of a database at an early stage of a sequencing project. The schematic is divided into three boxes arranged above one another. The top box shows a series of horizontal lines, each representing an individual contig. There are 35 contigs, and the length of the lines are proportional to the contig lengths. The line representing one of the two longer contigs is bisected by a vertical line which indicates it has been selected by the user. The next box below contains a display of all of this particular contig's individual readings. Each reading is represented by a line proportional to its length and positioned according to its place in the contig. Those above the midline are in their original orientation, those below have been complemented.

In the next box is a display of the quality plot for the contig above. At any point along the contig, the further the horizontal lines are from the midpoint, the lower the quality of the data. The top of the plot has height 2, and the base height −2. Lines at heights 1 or −1 indicate sections that are only well determined on one strand or the other. Lines at distances 2 or −2 indicate sections that are well determined on each strand but the two strands disagree. Where there is only a midline the sequence is well determined on both strands and they agree. So, for example, near the centre of this plot is a line stretching from −2 to 2 showing a section where the strands are well determined but do not agree. Just to the left of this point is a section containing a 0 to 1 line closely followed by a 0 to −1 line, indicating problems on alternate strands. At the very left of the display is a rectangle showing a section that is only well determined on one strand. This is confirmed by reference to the plot above which shows that this section is only covered by a single reading, and that this is in its original orientation.

that the local aligned sequences be shown in a text window. In this way the user can instantly ascertain the reason for a nonzero point in the quality plot, or can find the name of a reading in the plot of single contigs. To select a contig to display in the form of a gel reading schematic or quality plot, the user positions the crosshair on the relevant line in the top level display and clicks the mouse button.

The contig editor

The editor contains a set of function buttons and a text window showing the aligned sequences for an 80 character section of a contig. See figure 2. This text window is positioned within the contig using the scroll bar and scroll buttons. The scroll bar is located just above the window. Its length represents the length of the contig, while the shaded region represents the location and relative size of the currently visible portion. If mouse buttons are pressed when the mouse pointer is over the scroll bar, the window will be repositioned. The left mouse button causes the window position to move forwards, or right, one screen width. The middle mouse button repositions the window at a position along the contig corresponding to the point probed along the scroll bar. The right mouse button causes the window position to move


a
Contig Editor (V1.1)
Insert
Replace
Super Edit
Undo
Next Problem
Leave Editor
<<
<
>
>>
b
quit
hide
a34a1.s1RES
Scale up
Scale down
quit
G T A G A C A A C A T C T G A A A A A G G A G A C G G G G G A T C C T C T A G A G
a33e10.s1RES
Scale up
Scale down
quit
G T A G C A A C A A T C T G A A A A A G G A G A C G T G A A A G A G A A G A
a33g3.s1RES
Scale up
Scale down
quit
G T A G A C A A C A T C T G A A A A A G G A G A C G T G A A A G A G A A G A G
a33h11.s1RES
Scale up
Scale down
quit
G T A G A C A A C A T C T G A A A A A G G A G A C G T G A A A G A G A A G A G A

```
                     1170      1180      1190      1200      1210      1220      1230      1240
   -55 a33f2.s1      TGTCG
  -107 a34a1.s1      TGTCGTAGACAAC*ATCTGAAAAAGGAGACG
    51 a33e10.s1     TGTCGTAG*CAACAATCTGAAAAAGGAGACGTGAAAGAGAAGAGAGACACAGAGAAATGAGAAGACTTGGGTAATGAGTG
   -67 a33g3.s1      TGTCGTAGACAAC*ATCTGAAAAAGGAGACGTGAAAGAGAAGAGAGACACAGAGAAATGAGAAGACTTGGATAATGAGTG
   -86 a33h11.s1     TGTCGTAGACAAC*ATCTGAAAAAGGAGACGTGAAAGAGAAGAGAGACACAGAGAAATGAGAAGACTTGGATAATGAGTG
    93 a35a7.s1                                                                                   GTG
       CONSENSUS     TGTCGTAGACAAC-ATCTGAAAAAGGAGACGTGAAAGAGAAGAGAGACACAGAGAAATGAGAAGACTTGG-TAATGAGTG
```

Figure 2 shows a typical display from the contig editor. It is divided into two. Part 2a is a text window showing a section of a contig and part 2b shows a graphics window containing 4 traces from machine derived readings. The text window shows an 80 character segment of the contig which includes data from 6 readings. The names and numbers of the readings are shown in the left hand panel, minus signs indicating those which are the complements of their original. Along the top edge of the text window the contig positions are numbered in 10's, and at the bottom is a line containing the consensus for readings aligned above. Immediately above the main text window is a scroll bar which can be used for large movements along the contig, and which indicates the current window position within the contig. At the left of the scroll bar are the arrow buttons that provide for finer movements. At the top are the buttons for controlling the editing operations described elsewhere in the text. The segment of sequence shown in the editor window corresponds to the midpoint of the contig depicted schematically in figure 1. At position 1237 the section in figure 1 giving rise in the quality plot to a line from −2 to 2, is seen to be due to a definite G on one strand, and a definite A on the other. The padding symbols (*) at 1180 and 1175, which have been placed by the assembly routine, also account for the other problems (a typical compression pattern) referred to in the legend for figure 1.

The graphics window shown in 2b contains the traces for the readings covering the current cursor position (1190) in the contig editor. They are selected by clicking on the mouse button, and will appear instantly. On a colour screen the four bases A,C,G,T are shown by the colours green, purple, black and red, but in this figure, taken from a monochrome display, the bases are identified by four different line styles. Above each trace is a copy of the original sequence as interpreted by the sequencing instrument. Above this is a scroll bar that permits independent movement of each trace. Each trace is identified at the top left hand corner and underneath these names are separate buttons to enable changes in the way each individual plot is viewed. Note that the trace for reading a34a1.s1 includes the primer site, and hence the sequence has been clipped in the contig editor display.

backwards, or left, one screen width. Four scroll buttons allow the position of the window to be finely adjusted. Within the window a cursor indicates the point where edits will occur. The cursor can be positioned on a base by pointing the mouse over the base and pressing the left mouse button. The cursor can also be moved using the four arrow keys. The left and right arrows move the cursor forwards and backwards along the sequence. The up and down arrows move the cursor between the sequences. If, when using the arrow keys, the cursor moves off the screen, the window is repositioned with the cursor at the centre.

Insertions, deletions and replacements can all be performed using the editor. The editor operates in either 'Insert' or 'Replace' modes, and this is indicated by which of the two buttons at the top of the window is highlighted. The user toggles between the modes by clicking on the appropriate button. Edits can be performed on an individual aligned reading or on the consensus. Performing edits on the consensus is an efficient way to apply the same operation to all sequences aligned at the cursor position. Further, insertions and deletions on the consensus will also maintain the alignment of all sequences to the right of the cursor position, by moving them right and left respectively. As insertions and deletions in an individual reading can result in misalignments, these operations are prohibited under normal operation. A special button 'Super Edit' overrides this.

Bases can be changed when the editor is in replace mode. A character typed from the keyboard replaces the character at the cursor position. The cursor is then advanced one base. Bases can be inserted into a sequence when the editor is in insert mode. A character typed from the keyboard is inserted to the left of the character at the cursor position. Bases can be deleted when the editor is in any mode. Pressing the delete key causes the character to the left of the cursor position to be deleted. As an edit is performed the window is updated to reflect the change and the effect it has on the visible part of the consensus. A file is kept to record all edits. The previous edit can be negated using the 'Undo' button. A function 'Find Next Problem' helps identify areas where problems exist by moving the cursor to the position of the next unresolved base in the consensus. The changes are made to the database only when the editor is exited. Until then, all changes are made to a copy of the contig.

To resolve alignment problems the machine readable traces obtained from the ABI 373A and the Pharmacia A.L.F. can be displayed while the sequences are being edited. See figure 2b, and note that both types of trace and also conventional data can be mixed in a single project. By double-clicking with the mouse at the point in the edit window where a problem exists, the relevant part of the corresponding machine trace is displayed in a separate window. Double clicking on the consensus will automatically display several of the corresponding traces. Because of the accuracy with which individual traces are positioned, multiple traces will be precisely aligned. To conserve screen space, a maximum of four traces can be displayed at once, and when a further trace is requested, the top one in the window is automatically discarded. The user can discard any particular trace by pressing the 'Quit' button alongside the trace. Alternatively, all traces can be discarded by pressing the 'Quit' button at the top of the trace window. The individual traces can be rescaled or scrolled independently.

A further version of the editor allows two contigs to be joined. It is required for performing joins that the auto assembly function reports but which it cannot manage. From the auto assembly run the user will know the position of the overlap. The join editor allows the two contigs to be slid past one another until they are in register. At this point the user selects a 'Lock' button which forces the two contigs to move together as a single entity for all subsequent changes of position. A narrow window sandwiched between the displays of the two contigs shows disagreements between their consensus sequences. All the functions described for the contig editor are available from the join editor. The join is completed when the editor is exited.

The editor also has a facility to allow annotation (or tagging) of sequences. The types of annotations made can be tailored to the project being undertaken, but typically it might include marking the positions of primers used for walking, or sites such as compressions that have caused problems during sequencing. Each annotation of a particular type is highlighted in a specific colour. The section of sequence to annotate is selected using a standard mechanism for cutting and pasting, and then the 'Create Tag' function invokes a simple editor to enable the user to specify the annotation. The type of the annotation is selected from a menu of available alternatives and a comment may also be added by typing into the editor's text window. On leaving the editor the tag will be created and highlighted in the colour for the type chosen. Previous annotations can be modified by selecting the 'Edit Tag' function, invoking the same editor described before.

Hardware and software requirements

We have been developing the program on Sun SPARCstations under SunOS 4.1. To get the most from the program a colour screen is needed although it will still work on a monochrome system. The programs are written in FORTRAN77 and ANSI C and use X11 Release 4. Also see the discussion.

RESULTS AND DISCUSSION

We have described the current state of our program xdap, placing emphasis on the new contig editor. The constraint of trying to produce portable programs restricted the ease of use of the editors in previous versions, but X windows has provided the means of writing portable programs with convenient user interfaces. The new editor is easy to learn and to use, and scrolls rapidly along cosmid sized contigs.

The porting of the whole package from a version running on a Vax under the VMS operating system to one using X on the Sun has been described previously(7). So far we have only used the program on Sun workstations but we would expect it, and the other programs in the package, to work on most UNIX machines, or with a little more effort, other machines with X windows but different operating systems. We also have a version that does not require an X server, but which has all the facilities described except the new contig editor.

Although the program is in constant use by members of the C. elegans sequencing project, it is still under intensive development (the name is short for 'developing assembly program') and below we outline some of the anticipated improvements.

The graphical displays are still constrained by their tektronix origins and will be rewritten so that they operate in unison with the contig editor. Colour will be used. At present we are using a primer selection program, osp(8) which is separate from xdap, but we expect to incorporate this search procedure into the contig editor. Above we mentioned the use of the program ted for visual selection of the ends of machine read data. Initially this program was used because it also allows the sequences to be edited.

However as we no longer edit the sequences at this stage we will replace it by a program that chooses the cutoff points without user intervention.

The C. elegans project(5) is using a combined shotgun and primer walking strategy to sequence cosmid sized segments of the genome. Several cosmids are being sequenced in parallel, with, so far, one of length 41kb finished and several others nearing completion. To date, in the Cambridge laboratory alone, more than 625kb of raw data have been handled by xdap. Obviously the success rate for automatic assembly depends on the quality of the data and the frequency and extent of repeats, but we estimate that better than 95% of readings have been assembled without user intervention. The contig editor, combined with the graphics displays and the facility to view traces instantly has substantially reduced the amount of time spent editing the data. They have also shortened the time required for planning new experiments to double strand the sequence, join contigs or resolve compressions. The planned improvements should make for further reductions in user time.

Although the problems tackled by xdap are important and topical we are aware of no publications on the subject for several years(9,10). However we believe some more recent unpublished methods are available, but only in commercial packages.

A record is kept of all changes to the individual sequence readings. This permits precise alignment of traces and sequences, and allows the user to recall the original data. It is a further tool to aid the essential task of ensuring the accuracy of the final sequence. We would like to point out that it also provides a method for systematic monitoring of the types of base calling errors made by machines and people, and such information could be valuable for improving the base calling software.

ACKNOWLEDGEMENTS

We thank all members of the C. elegans sequencing project for their ideas and enthusiasm. For contributions to the software thanks are due to Richard Durbin, Tim Gleeson, LaDeana Hillier and John Sulston. We also acknowledge the cooperation of ABI and Pharmacia in agreeing to our routines accessing the output files of the 373A and the A.L.F.

REFERENCES

1. Staden,R. (1982). Nucl. Acids Res. 10, 4731–4751
2. Staden,R. (1990). Comp. Applic. Biosc. 6, 387–393.
3. Baer,R., Bankier,A.T., Biggin,M.D., Deininger,P.L., Farrell,P.J., Gibson,T.J., Hatfull,G., Hudson,G.S., Satchwell,S.C., Seguin,C., Tuffnell,P.S. and Barrell,B.G. (1984). Nature 310, 207–211.
4. Chee,M.S., Bankier,A.T., Beck,S., Bohni,R., Brown,C.,M., Cerny,R., Horsnell,T., Hutchison III,C.A., Kouzarides,T., Martignetti,J.A., Satchwell,S.C., Tomlinson,P., Weston,K.M. and Barrell,B.G. (1990) Current Topics in Microbiology & immunology 154, 125–169.
5. Roberts,L (1990) Science 248, 1310–1313.
6. Hillier,L and Gleeson,T.J (in preparation)
7. Gleeson,T.J and Staden,R. Comp. Applic. Biosc. In press.
8. Hillier,L and Green,P (in preparation)
9. Peltola,H. Soderland,H. and Ukkonen,E. (1984) Nucl. Acids Res. 12, 307–321.
10. Johnson,R.E., Mackenzie,J.M.Jr. and Dougherty,W.G. (1986). Nucl. Acids Res. 14, 517–527.