Nucleic Acids Research, Vol. 19, No. 17 4663–4667

Software tools for analyzing pairwise alignments of long sequences

Scott Schwartz, Webb Miller*, Cher-Ming Yang and Ross C.Hardison[1]
Department of Computer Science and [1]Department of Molecular and Cell Biology, The Pennsylvania State University, University Park, PA 16802, USA

Received May 29, 1991; Revised and Accepted August 2, 1991

ABSTRACT

Pairwise comparison of long stretches of genomic DNA sequence can identify regions conserved across species, which often indicate functional significance. However, the novel insights frequently must be winnowed from a flood of information; for instance, running an alignment program on two 50-kilobase sequences might yield over a hundred pages of alignments. Direct inspection of such a volume of printed output is infeasible, or at best highly undesirable, and computer tools are needed to summarize the information, to assist in its analysis, and to report the findings. This paper describes two such software tools. One tool prepares publication-quality pictorial representations of alignments, while another facilitates interactive browsing of pairwise alignment data. Their effectiveness is illustrated by comparing the β-like globin gene clusters between humans and rabbits. A second example compares the chloroplast genomes of tobacco and liverwort.

INTRODUCTION

Alignments of DNA or protein sequences are very informative for a variety of life sciences. Two sequences may have a number of similar regions, and it is desirable to obtain all of the significant local alignments. Local alignments match individual regions within larger sequences, whereas global alignments force a match throughout the two sequences being compared. Gap-free local alignments have a readily computed statistical significance associated with them (1), but they are, of course, generally shorter than alignments containing gaps. Gap-free alignments of very long sequences can be generated quickly using the program BLAST (2). However, alignments containing gaps are often more informative, especially when the spacing between conserved regions is apparently not critical. Several alignment programs based on the original algorithm of Smith and Waterman (3) are in current use. For example, LFASTA (4, 5) is fairly rapid, and SIM (6) is efficient in its use of computer space.

After finding all the significant local alignments between two long sequences, one is frequently left with such a large amount of information that adequate analysis may seem impossible. We have developed two software tools that enable the user to better comprehend alignment data. The first tool, called LAD (Local Alignment Diagramer), displays alignments along with known features of the sequences, such as exons and repetitive elements, so that correlations between alignments and sequences features can be easily visualized. The second tool, called LAV (Local Alignment Viewer), supports interactive browsing of alignments. One can begin with a display showing all the computed alignments, examine the alignments at varying stringencies, zoom in on interesting regions, and directly inspect underlying alignments.

METHODS

The human β-like globin gene cluster is from Collins and Weissman (7; GenBank entry HUMHBB) and the rabbit β-like globin gene cluster is from Margot *et al.* (8; GenBank entry RABBGLOB). The tobacco chloroplast genome sequence is from Shinozaki *et al.* (9; Genbank entry TOBCPCG) and the liverwort chloroplast genome sequence is from Ohyama *et al.* (10; Genbank entry MPOCPCG).

LAD and LAV were implemented in the programming language C++. LAD generates PostScript output, and the LAV user interface utilizes the InterViews toolkit (11) under X Windows. All software was written and used on Sun workstations running the Unix operating system.

RESULTS

LAD

Figure 1 is the LAD (Local Alignment Diagramer) plot of 195 local alignments between a 73 kb sequence containing the human β-like globin gene cluster and a homologous 44 kb sequence of rabbit, as computed by the SIM program (6). Alignments are displayed by analogy with a traditional dot-plot (12,13), as was done earlier by the PLFASTA alignment program of Pearson and Lipman (4,5).

Figure 1 (minus several hand-drawn arrows) was produced by LAD directly from the SIM alignments and from hand-edited

* To whom correspondence should be addressed

files specifying feature locations and their pictorial representations. To produce a plot of this sort, one first runs an alignment program (e.g., SIM or BLAST) and collects the output in a file. The alignment and features files are processed by the LAD program, and then sent to a PostScript printer.

A LAD-generated plot provides an overview of the alignments in a compact, readily understood format. For instance, Figure 1 shows extensive matches between orthologous globin genes (e.g., between the ϵ gene in each species), extending through exons, introns, and several kilobases of flanking DNA. Matches between paralogous globin genes (e.g., between rabbit ϵ and human $^{G}\gamma$) are generally limited to the three exons of each gene (the nonmatching small introns are not visible at this resolution). Matches between the coding regions of L1 repeats are also apparent. Many of these features were previously seen in an extensive dot-plot analysis of these same data (8), and in fact the initial reaction to viewing output from LAD is that one has a very clean dot-plot. A principal advantage of LAD is that each line represents a specific alignment that can be directly accessed by the user. A second advantage is that the symbols depicting sequence features, located on the plot's borders, are placed exactly and objectively. This avoids the inaccuracy, and the susceptibility to personal bias, of hand-drawn symbols.

The use of LAD to display alignments, especially SIM alignments, regularly reveals features that were not detected by dot-plot analysis. Although the presentations of a dot-plot and a LAD plot are similar, the underlying concepts are quite different. In essence, a conventional dot-plot shows all gap-free alignments of a simple form that exceed some scoring threshold. It does not have the capacity to detect a chain of gap-free alignments based on their cumulative score. Alignments attempt to extend regions of similarity, whereas a dot-plot will extend a collection of dots in a non-meaningful direction. Thus, dot-plots miss faint but significant alignments that contain frequent gaps; if the cutoff score is set low enough to be satisfied by the gap-free segments of the alignment, then there will be so many matches that the alignment will be lost in background noise. The BLAST program suffers from the same inability to handle gaps.

As a second example of a LAD plot, Figure 2 shows alignments of the tobacco and liverwort chloroplast genomes computed by the BLAST program. These genomes are two to three times larger than the mammalian globin gene clusters in the previous example. As expected (14), matches throughout much of the genome are apparent. The segment of the chloroplast genome from *rpoB* to *psbA* is inverted in liverwort relative to tobacco (hence the diagonal with downward slope). The rRNA genes are duplicated by an inverted repeat, producing the symmetrical X pattern in the upper right quadrant of Figure 2. The off-diagonal matches in the *psaA-psaB* region show that these genes encoding the two large subunits of photosystem I resulted

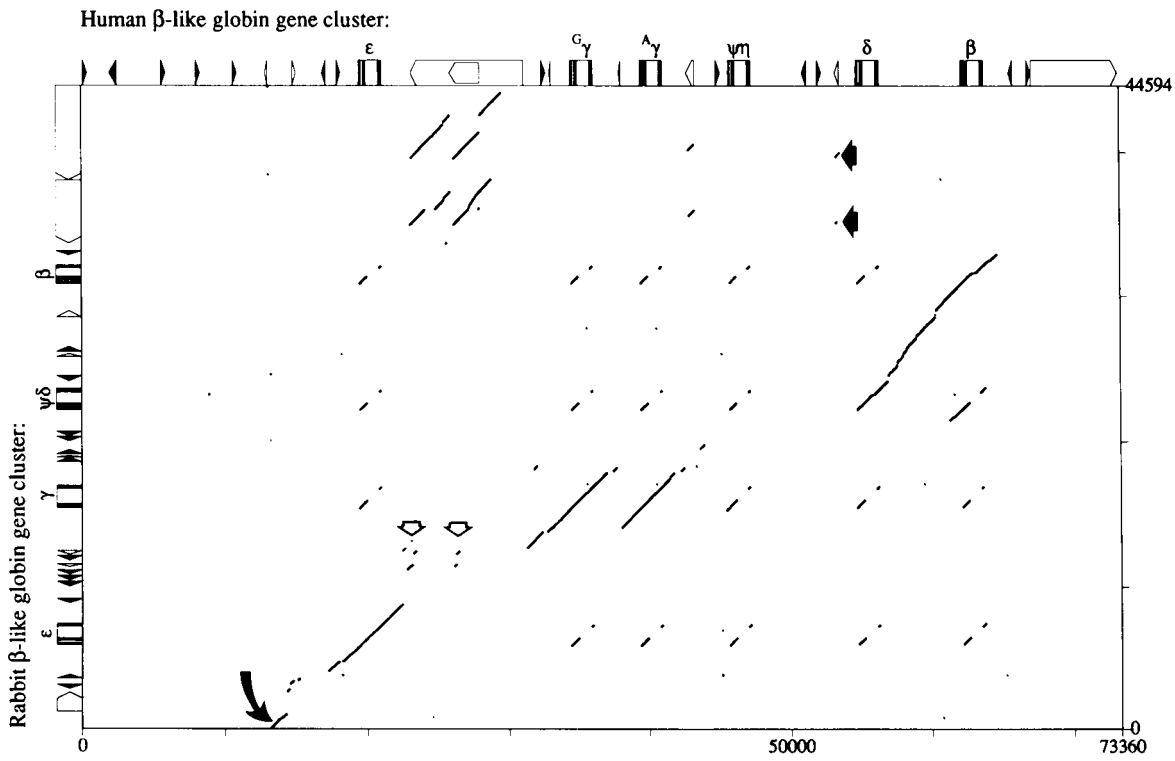

Figure 1. Plot produced by the LAD program of the 195 highest-scoring local alignments between a 73360-nucleotide sequence from the human β-like globin gene cluster and an analogous 44594-nucleotide sequence of rabbit DNA. Alignments were computed by the program SIM (6) with a score of 1 for matches, -1 for mismatches, -4 for opening a gap, and -0.4 for each symbol in the gap. The computation took approximately 15 hours on a Sun SparcStation1+ workstation. Alignments are displayed only if their score exceeds a threshold $\tau = 23$, chosen so that the probability is 0.05 that random sequences matching the given sequences in length and nucleotide composition have a gap-free alignment scoring at least τ. Genes are shown as boxes with exons filled and introns unfilled. *Alu* (human) and C (rabbit) repetitive elements are indicated by black triangles; L1 repeats are represented as unfilled boxes with a pointed end. A few arrows have been added by hand to indicate features discussed in the text. The curved arrow at the lower left indicates the alignment shown in Figure 3. Open arrows point to alignments that reveal the existence of truncated L1 repeats in the 3′ flank of the rabbit ϵ gene. Solid arrows indicate matches revealing the presence of a previously-undetected L1 repeat in the 5′ flank of the human δ gene (sequence locations 53274–53592).

from a duplication. The other major alignment off the main diagonal involves the largest gene in each of the two chloroplast genomes. The 3′ end of tobacco ORF2280 aligns in several segments with the 3′ end of liverwort ORF2136. Most of the remainder of the ORFs do not match, and the ORFs are located in different positions of the genome.

LAV

LAV (Local Alignment Viewer) is a second tool for managing pairwise sequence alignments. When first invoked, LAV presents the user with a LAD-like representation of the alignments. One can focus in on smaller regions of interest by drawing a box around the target region using the mouse or call up a standard, nucleotide-resolution representation of any chosen alignment as another window on the screen of the workstation (Figure 3).

This approach avoids a major problem that arises when a traditional dot-plot program is used together with an unrelated alignment program. The dot-plot may suggest matches that are not detected by the alignment program, and vice versa. In contrast, features visible in the pictorial representation presented by LAV are guaranteed to be available at full alignment resolution.

LAV has several additional features. The pictured alignments can be viewed at any desired stringency by altering the number of alignments viewed or (for BLAST alignments only) by setting a threshold for statistical significance. The threshold similarity score for viewing subsegments of a SIM alignment (i.e., between gaps) can also be varied. The sequential viewing of alignments at increasing similarity scores or significance values is analogous to conducting a 'melting experiment' directly on the screen, and an appropriate method for scoring mismatches could even account for differences in thermal stability between GC versus AT base pairs.

In addition, there is a facility to simultaneously display two sets of alignments. This capability is quite useful for comparing the alignments generated by different algorithms or by the same algorithm with different parameters. Another use is for highlighting portions of alignments where both of the matched

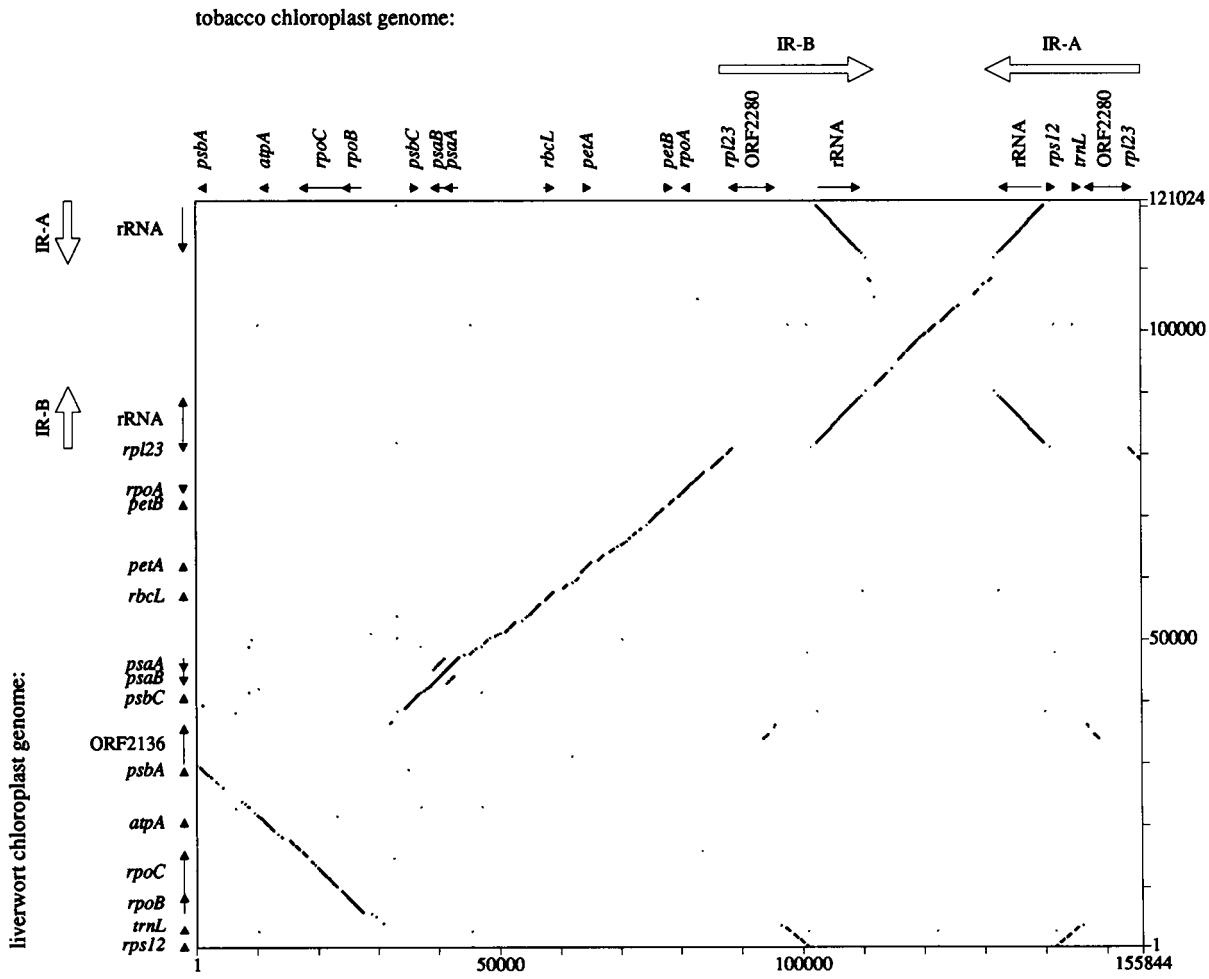

Figure 2. Plot produced by the LAD program of the 253 highest-scoring gap-free local alignments between the chloroplast genomes of *Nicotiana tabacum* (tobacco) and *Marchantia polymorpha* (liverwort). Lines running from lower left to upper right plot the alignments generated between the two sequence files as listed in GenBank; the origin is at the lower left corner. Lines slanted downward plot the alignments between the reverse complement of the liverwort chloroplast sequence and the standard orientation of the tobacco chloroplast sequence; the origin is at the upper left corner. Alignments were computed by the BLAST program (2) with a score of 1 for matches and −1 for mismatches. A gap-free alignment is depicted in this plot only if (*i*) it contains an exact match of 8 consecutive nucleotides and (*ii*) its score exceeds the threshold $\tau = 27$, chosen so that the probability is 0.05 that random sequences matching the given sequences in length and nucleotide composition have a gap-free alignment scoring at least τ. The computation required about 60 seconds on a Sun SparcStation2 workstation. The positions and orientations of selected genes and the inverted repeats (IR-A and IR-B) are shown to provide landmarks along the genome.

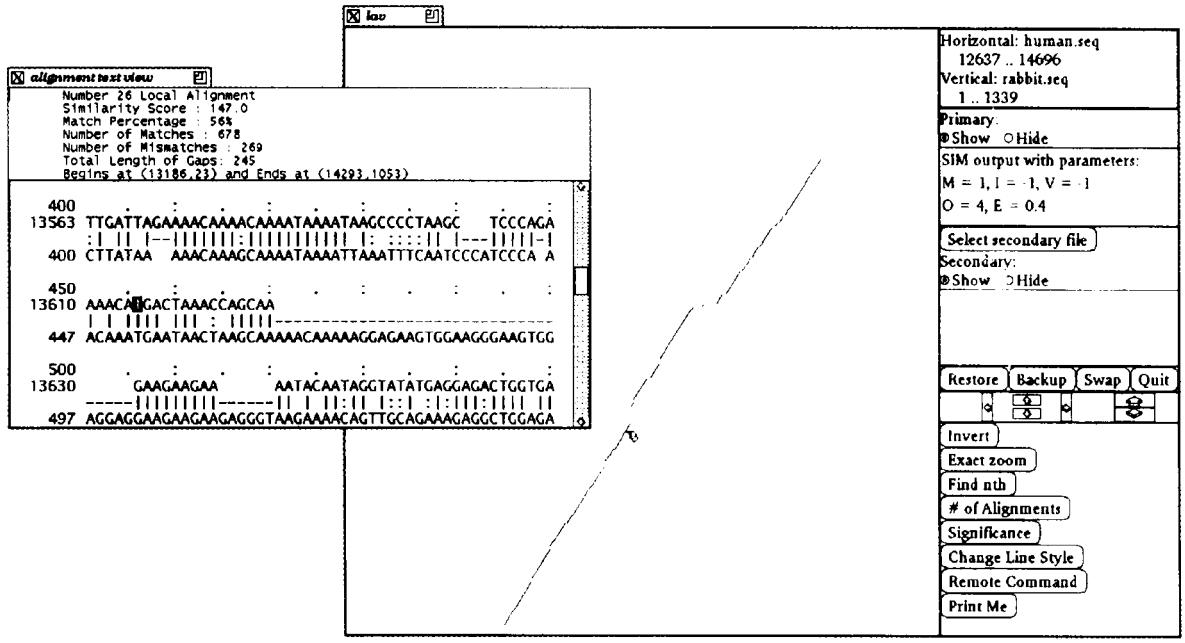

Figure 3. Contents of the computer screen after (*i*) invoking LAV to browse the alignments pictured in Figure 1, (*ii*) zooming in on the alignment pointed to by the curved arrow in Figure 1 and (*iii*) clicking the middle mouse button. The selected alignment is displayed, with the first character of the aligned pair nearest the mouse location shown in reverse type.

regions exhibit some interesting sequence feature, such as an open reading frame or a certain regulatory signal. These matches might be automatically detected by a program that reads the alignments.

DISCUSSION

We have developed two new programs to aid in the analysis and interpretation of the voluminous output generated when local alignments are computed between two large sequence files. LAD generates a graphical display of all alignments, with key sequence features plotted objectively along the axes, while LAV is interactive with the user and is highly versatile.

Several earlier software packages, such as the IBI-Pustell programs, have some of the annotated dot-plot capabilities of LAD/LAV. Among the advantages of LAD/LAV are: (1) use with SIM alignments allows gaps (which are sometimes necessary to detect subtle relationships), (2) use with BLAST alignments provides a well-founded estimation of statistical significance, (3) use with either BLAST or SIM alignments avoids some of the sensitivity to parameter/threshold settings for which dot-plots are notorious, (4) LAD/LAV takes advantage of modern hardware and software (e.g., PostScript printers and the X Windows software package), and (5) the software is freely available.

Some new L1 repeats were discovered when SIM + LAD was used to compare the β-like globin gene clusters of rabbits and humans. In both species, the region between ϵ and γ has been the site of multiple, recursive integrations of repeats, which leads to some ambiguity in inferring the evolutionary history. In particular, several alignments indicate that the L1 repeats between ϵ and γ inserted separately in rabbit and human, but one alignment indicates that the L1s have a common flanking sequence, implying that these L1 repeats are derived from an L1 sequence in the last common ancestor. If the latter interpretation is correct, this is the only known example of an L1 repeat in a common position in the genomes of species from different mammalian orders (8,15).

The newly-discovered L1 repeats in the rabbit (between ϵ and γ) and human (5′ to δ) β-like globin gene clusters are unusual in two respects. First, they align as well or better between species than within species. The alignments range between 58% and 70% matches in both types of alignments, whereas most L1s show $> 90\%$ similarity within a species but only about 65% similarity between species (16). The greater similarity of LIs within a species most likely reflects their recent amplification by transposition, with possibly some contribution from concerted evolution (17). Second, they do not contain sequences from the 3′ untranslated regions that follow the long ORF2 of L1 repeats (17). Both of these characteristics support the hypothesis that these are ancient, highly divergent L1 repeats, distinct from the more homogeneous, well-studied families of L1 repeats that are still active in transposition (18). The greater divergence of these older L1s make them difficult to detect in conventional dot-plots. Perhaps these more divergent L1s represent precursors to the more commonly studied L1 repeats. These latter L1s may have acquired their distinctive 3′ untranslated regions by fusion with the ancestral coding sequences (15). The highly divergent L1s identified here may be analogous to the ancestral rodent L1 discussed by Pascale *et al.* (19).

Development of our software for generating and analyzing pairwise alignments is continuing. One project is to write a utility program that will automatically generate files containing sequence feature information, such as that used to decorate the borders of Figure 1, from GenBank features tables. Also, we are developing programs to 'merge' several pairwise alignments for simultaneous comparison of three or more sequences. Information gathered by this approach complements that obtained from traditional multiple-alignment programs.

AVAILABILITY

The source code for LAD and LAV, together with versions of SIM and BLAST suitable for use with them, is available by anonymous ftp from groucho.cs.psu.edu. SIM and BLAST are written in C and are portable to a wide range of machines. BLAST has two versions, one for DNA sequences and one for proteins, while SIM handles both kinds of sequences. LAD and LAV require C++ compilers. LAD output can be printed on a PostScript device. LAV should be reasonably portable to machines supporting the InterViews toolkit (11). The current implementation of InterViews runs on any Unix system with X Windows, and ports to other systems are underway. Information about InterViews can be obtained by sending electronic mail to interviews-request@interviews.stanford.edu. The Unix implementation, documentation, and descriptive papers can be obtained by Unix users with access to the Internet via anonymous ftp from interviews.stanford.edu.

The authors can be contacted by electronic mail at webb@cs.psu.edu.

ACKNOWLEDGMENTS

We thank Bill Pearson for graciously providing many useful comments on a preliminary draft of the paper. Mark Boguski helped test the software on protein sequences, and Jim Ostell explained the capabilities of his IBI/Pustell program for dot-plots. S.S., W.M. and C.-M.Y. were supported in part by grant R01 LM05110 from the National Library of Medicine. R.C.H. was supported by PHS grant RO1 DK27635 and an RCDA DK01589.

REFERENCES

1. Karlin,S. and Altschul,S.F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
2. Altschul,S.F., Gish,W., Miller,W., Myers,E.W and Lipman,D.J. (1990) *J. Mol. Biol.* **215**, 403–410.
3. Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol.* **147**, 195–197.
4. Pearson,W.R. and Lipman, D. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
5. Pearson,W.R. (1990) *Methods Enzymol.* **183**, 62–98.
6. Huang,X., Hardison,R.C. and Miller,W. (1990) *Computer Applications in the Biosciences* **6**, 373–381.
7. Collins,F.S. and Weissman,S.M. (1984) *Progr. Nucl. Acids Res. Mol. Biol.* **31**, 315–462.
8. Margot,J.B., Demers,G.W. and Hardison,R.C. (1989) *J. Mol. Biol.* **205**, 15–40.
9. Shinozaki,K., Ohme,M., Tanaka,M., Wakasugi,T., Hayashida,N., Matsubayashi,T., Zaida,N., Chunwongse,J., Obokata,J., Yamaguchi-Shinozaki,K., Ohto,C., Torazawa,K., Meng,B.Y., Sugita,M., Deno,H., Kamogashira,T., Yamada,K., Kusida,J., Takaiwa,F., Kato,A., Tohdoh,N., Shimada,H., Sugiura,M. (1986) *EMBO J.* **5**, 2043–2049.
10. Ohyama,K., Fukuzawa,H., Kohchi,T., Shirai,H., Sano,T., Sano,S., Umesono,K., Shiki,Y., Takeuchi,M., Chang,Z., Aota,S.-I., Inokuchi,H. and Ozeki,H. (1986) *Nature* **322**, 572–574.
11. Linton,M. A., Vlissides,J.M. and Calder,P.R. (1989) *Computer* **22**, 8–22.
12. Maizel,J.V. and Lenk,R.P. (1981) *Proc. Natl. Acad. Sci. USA* **83**, 7665–7669.
13. Pustell,J. and Kafatos,F.C. (1982) *Nucleic Acids Research* **10**, 4765–4782.
14. Palmer,J.D. (1991) In *The Molecular Biology of Plastids*, L. Bogorad and I. K. Vasil, eds., Vol. 7 in *Cell Cultures and Somatic Cell Genetics in Plants* (I. K. Vasil, Ed.-in-Chief). In press.
15. Demers,G.W., Matunis,M.J. and Hardison,R.C. (1989) *J. Mol. Evol.* **29**, 3–19.
16. Scott,A.F., Schmeckpeper,B.J., Abdelrajik,M., Comey,C.T., O'Hara,B., Rossiter,J.P., Cooley,T., Heath,P., Smith,K.D. and Margolet,L. (1987) *Genomics* **1**, 113–125.
17. Weiner,A.M., Deininger,P.L. and Efstratiadis,A. (1986) *Annu. Rev. Biochem* **55**, 631–661.
18. Kazazian,Jr.,H.H., Wong,C., Youssoufian,H., Scott,A.F., Phillips,D.G. and Antonarakis,S.E. (1988) *Nature* **332**, 164–166.
19. Pascale,E., Valle,E. and Furano,A.V. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 9481–9485.