Macintosh Sequence Analysis Software

DNAStar's LaserGene

Jonathan P. Clewley

Abstract

The analysis of information in nucleotide and amino acid sequence data from an investigator's own laboratory, or from the ever-growing worldwide databases, is critically dependent on well planned and written software. Although the most powerful packages previously have been confined to workstations, there has been a dramatic increase over the last few years in the sophistication of the programs available for personal computers, as the speed and power of these have increased. A wide choice of software is available for the Macintosh, including the LaserGene suite of programs from DNAStar. This review assesses the strengths and weaknesses of LaserGene and concludes that it provides a useful and comprehensive range of sequence analysis tools.

Index Entries: DNA and protein sequence analysis; DNAStar; LaserGene; GeneJockey.

1. Introduction

There are several comprehensive (and expensive) DNA and protein sequence analysis packages available for the Apple Macintosh: DNAStar's LaserGene, MacVector *(1)*, GeneWorks, MacDNAsisPro, and MacMolly Tetra. Mangalam referred to the first four of these programs as the Gang of Four *(2)*—if MacMolly Tetra is included with them they should be called the Gang of Five, referred to herein as G5. There are also less expensive packages with (slightly) fewer features, such as DNAStrider *(3)* and GeneJockey, which has many of the facilities of the G5. More specialized programs exist for oligonucleotide analysis, secondary structure prediction, multiple alignment, and gene (plasmid) design, some of which are in the public domain or are shareware. Should you consider buying one of the G5, one of the medium sized programs, use GCG online *(4)*, or should you collect public domain programs? Some of these programs were reviewed in 1991 *(5)*; there may be more recent information available by anonymous ftp from ftp.bio.indiana.edu in the directory /molbio/mac/pm-macinmolbio.txt.

Unlike MacVector, for example, LaserGene is a suite of programs rather than one single module. It has a menu driven interface (the LaserGene Navigator) that allows selection between the different modules; alternatively they may be started by double-clicking on their icons. This module approach has been criticized as a weak feature of LaserGene *(6)*. The company would no doubt argue that it allows users to purchase only those modules that are appropriate to their needs. What is more important, it allows easier addition of new programs, which is perhaps its greatest strength. At present there are seven modules: sequence editing and analysis, restriction analysis and mapping, multiple sequence alignment, sequencing

*Address to which all correspondence and reprint requests should be sent: Virus Reference Division, Central Public Health Laboratory, 61 Colindale Avenue, London NW9 5HT, UK.

project management, biological database resource, protein analysis, and primer selection.

2. Sequence Editing and Analysis (EditSeq)

The usual starting point for use of LaserGene will be the creation of sequence files: These may be either keyed in or imported from a variety of other formats. Often, they will be files retrieved from the databases supplied on CD-ROM (*see* GeneMan, later). The sequence is displayed in the upper half of a scrollable window, information (comments) about it in the lower half. Both DNA and protein sequences can be manipulated in several ways. They can be translated or (proteins) reverse-translated with a selected genetic code, proofread, changed to the other strand or to the reverse sequence, and so on. The weakest features of this module are the search for ORFs and the lack of export formats. When I need these features, I turn to GeneJockey. Although it is possible to find all the ORFs in a DNA sequence with EditSeq, it lacks the visual clarity and ease with which this may be achieved with GeneJockey. This is understandable, as I think it reflects how the problem was solved by the programmers. What is less fathomable is the inability of EditSeq to export or save files in any format other than LaserGene or text files with the comments included. Commonly, one wants to prepare plain sequence files (ASCII) for transfer of the sequences to other programs (for example, PHYLIP, CLUSTALV, or MACAW) or other operating systems (DOS, UNIX, or VMS) without having to resort to cutting and pasting, as EditSeq forces one to do. GeneJockey, on the other hand, allows saving as a plain text file without comments that can be unambiguously read by other programs.

3. Restriction Analysis and Mapping (MapDraw)

Analysis and display of the restriction sites in a DNA sequence are accomplished very quickly using MapDraw. The default display is of the sequence of both strands of DNA and translation of each. Although translation of either or both strands can be turned off, it does not appear possible to display just the coding strand and one reading frame. Somewhat surprisingly, ORFs can be displayed from within MapDraw, rather than from within EditSeq, but they are not capable of being expanded into protein files by clicking, as can be done in GeneJockey. The results of the restriction mapping analysis can be printed out in a variety of useful formats.

4. Multiple Sequence Alignment (MegAlign)

Sequences can be aligned either in pairs or in multiples, and do not have to be the same length. The pairwise alignment of DNA either can be by a combination of the Martinez and Needleman-Wunsch approaches, or by the Wilbur-Lipman method. Guidelines are given for the appropriate use of either method, but it is often best to try both when searching for similarity between two uncharacterized sequences. Multiple alignment can be either by the Jotun Hein *(7)* or Clustal *(8)* methods. Protein alignment is by the Lipman-Pearson method. The multiple alignment capability is a strength of LaserGene, as large alignments can be accomplished on machines with relatively small amounts of RAM—provided you are prepared to wait.

5. Sequencing Project Management (SeqMan)

ABI sequence chromatograms can be imported into and viewed in SeqMan, an important feature if a laboratory is using this machine. The alternative to using SeqMan for this is to use ABI's own software or a more specialist program, such as Sequencher *(9)*. This laboratory has found that SeqMan is suitable for reasonable chromatographic data from the ABI machine, but that messy traces are better interpreted using SeqEd from ABI. Among the other G5 programs, none have (at the time of writing) the ability to import and display ABI traces. Of the other programs, GeneJockey II can do this and, helpfully for the red-green color blind, the color trace for the bases can be changed. SeqMan will also detect and remove vector sequences from the individual sequences that go into a contig.

6. Biological Database Resource (GeneMan)

DNAStar currently provides the EMBL and GenBank DNA sequence combined database, genetic Medline abstracts, Prosite, and the Swiss, PIR, and translated EMBL/GenBank protein databases on one CD. The Brookhaven protein databank (PDB files) is supplied on another CD. Previous releases have included the Los Alamos HIV database, the Berlin 4 and 5S RNA sequences, and the *E. coli* sequence database. Sequence records can be searched for by sequence or text, and files downloaded in EditSeq or text format. There is a link between the sequence and genetic Medline records, so that abstracts of papers associated with a selected sequence can be easily pulled out, if they are available. This is obviously not as sophisticated as Entrez *(10)*, but it is a useful feature. The Brookhaven files can be viewed as skeletons (α-carbon with or without side chains) or filled spheres (orbitals). The molecule can be rotated and viewed in stereo if the appropriate stereoviewer glasses are purchased. However, the PDB files cannot be directly exported from the CD for inspection or import into another program such as RasMol *(11)*.

7. Protein Analysis (Protean)

On opening a protein sequence in Protean, secondary structure predictions for it are displayed (Garnier-Robson, Chou-Fasman, Kyte-Doolittle, Eisenberg, Karplus-Schultz, Jameson-Wolf, Emini). The protein can also be shown as a linear space-filled model and as its chemical formula. Composition information and the titration curve are also available. It does not have the facility to predict proteolytic enzyme cutting sites, which is a feature of MacVector.

8. Primer Selection (PrimerSelect)

This is the newest module in the LaserGene suite of programs, and this review concerns version 0.93, which presumably will be updated progressively. PrimerSelect is the least intuitive program in LaserGene, and reference to the manual is essential. This is itself instructive about properties of successful primers, so it is not wasted time. The program offers a rapid selection of primers from either a single template or non-overlapping targets, and it ranks the primers found according to their likely amplification ability. The conditions employed by the program to select primers can be modified by the user. The physical characteristics of the primers and amplicon, as well as optimal annealing temperature for PCR, are calculated. It is fairly straightforward to use PrimerSelect for the design of nested primers. Degenerate, inosine, and backtranslated primers can all be found. Restriction sites can be engineered into the selected primers. Finally, the primer sequences can be printed out as an oligonucleotide synthesis request form. Altogether, the program feels powerful, but is awkward to use, perhaps reflecting the complexities of programming a comprehensive primer selection algorithm, as this is a criticism that can also be made of OLIGO. My personal wish list for upgrading PrimerSelect would be for it to accept multiple alignments, or to be able to work out primers capable of amplifying a group of related sequences, or to find primers that would amplify some specified sequences, but not prime from others. The only program I have found that comes close to achieving the latter is part of the B&L Utilities (Busch & Lucas Wissenschaftliche Software, Kunzenweg 22, D-7800 Freiberg i.Br., Germany). For example, this facility is useful for designing primers that would amplify all enteroviruses, but not the rhinoviruses. However, the real test of any primer selection program is how well the chosen primers work in PCR. The primers that PrimerSelect has devised for this laboratory are as yet insufficiently tried, so that test is still to come.

9. Conclusions

DNAStar's LaserGene is probably the best choice if buying a large DNA and protein sequence analysis program. It may lack the elegant interface appearance of MacVector, but its module approach allows the company to add new capabilities to the whole program suite in an effective way. After all, GCG is also a collection of programs. LaserGene's weaknesses are its ORF prediction, lack of export facilities, and, at present,

the lack of a DNA/RNA secondary structure prediction algorithm—DNAStar is apparently developing this. Secondary structure prediction is a necessity if sequencing rRNA genes. This laboratory *(12)* has used MUFOLD/loopDloop *(13)* and MacDNASIS for this purpose.

Any laboratory that is using an ABI sequencer and decides to buy one of the G5 programs will want one that imports ABI chromatograms, which limits the choice to LaserGene. Of the medium-cost and- sized programs, there is a choice between DNAStrider and GeneJockey. GeneJockey II will import ABI chromatograms and offers multiple alignment, so it now ranks with the "bigger" programs—the Gang of Six (G6?)—and offers a good compromise between features and cost.

The alternative to commercial programs is to use online facilities (GCG), if they are available locally. Even if they are, however, suitable Mac/PC programs will be required by most people doing a lot of sequence analysis, for their portable or home computer. Some of these are available as public domain programs, but, increasingly, the power and sophistication of the commercial programs make them more desirable. Of these, LaserGene is a good choice.

References

1. Olson, S. A. (1994) MacVector: an integrated sequence analysis package, in *Computer Analysis of Sequence Data, Part II* (Griffin, A. M. and Griffin, H. G., eds.), Humana, Totowa, NJ, pp. 195–201.
2. Mangalam, H. (1993) Striding the turf of the Gang of Four. *Trends Biochem. Sci.* **18,** 187,188.
3. Douglas, S. E. (1994) DNA Strider, in *Computer Analysis of Sequence Data, Part II* (Griffin, A. M. and Griffin, H. G., eds.), Humana, Totowa, NJ, pp. 181–194.
4. Griffin, A. M. and Griffin, H. G. (eds.) (1994) *Computer Analysis of Sequence Data, Part I,* Humana, Totowa, NJ.
5. Markiewicz, P. (1991) Computer software for molecular biology. *BioTechniques* **10,** 756–763.
6. Ahern, K. (1992) Macintosh sequence analysis software review: a comparison of functions. *Gen. Eng. News* **12,** 6,7.
7. Hein, J. J. (1990) Unified approach to alignment and phylogenies. *Meth. Enzymol.* **183,** 626–645.
8. Higgins, D. G. and Sharp, P. M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Computer Appl. Biosci.* **5,** 151–153.
9. Salser, W., Bedilion, T., Fitz-Gibbon, S., Mohajer, P., and Hansen, S. (1993) DNA sequence assembly and editing products which permit direct visulaization of raw data traces from automated (fluorescent) sequencing data. *J. NIH Res.* **5,** 81,82.
10. Cockerill, M. (1994) A versatile tool for retrieving molecular sequences: Entrez. *Trends Biochem. Sci.* **19,** 94,95.
11. Sayle, R. and Bissell, A. (1992) RasMol: A program for fast realistic rendering of molecular structures with shadows, in *10th Eurographics UK '92 Conference.* University of Edinburgh.
12. Linton, D., Clewley, J. P., Burnens, A. P., Owen, R. J., and Stanley, J. (1994) An intervening sequence (IVS) in the 16S rRNA gene of the eubacterium *Helicobacter canis. Nucleic Acids Res.* **22,** 1954–1958.
13. Gilbert, D. G. (1992) *Mulfold,* anonymous ftp to ftp.bio.indiana.edu.