TECHVIEW: SOFTWARE

All-in-One Sequence Analysis

Indira Rajagopal

When nucleic acid sequencing was in its infancy, the mere determination of the sequence of a stretch of DNA was news. Improvements in sequencing methods since that time have led to the rapid sequencing of many millions of base pairs from the genomes of a variety of organisms. The sequences that now fill databases have been determined not only by the research groups involved in the major genome sequencing projects, but also piecemeal by individual researchers investigating their genes of interest. Nowadays the main task is not sequence determination; it is sequence analysis, with the help of desktop programs.

MacVector 6.5
Oxford Molecular Group
$2950
$395/$795 (upgrade)
www.oxmol.com

Like the sequencing techniques themselves, the software available to carry out sequence analysis has become increasingly powerful and swift. The choices are varied, and the programs available range from those with a single function, such as the design of primers for polymerase chain reaction (PCR), to multipurpose products. MacVector is a "full-service" sequence analysis application that enables molecular biologists to analyze DNA and protein sequences on their Macintosh or PowerPC computers. The typical molecular biologist might want to enter sequence data obtained in the laboratory, then determine if it encoded any proteins or contained any special sequence motifs or restriction sites, construct maps that display any or all of the features of interest, and carry out sequence similarity searches against known sequences. If a nucleic acid sequence appears to encode a protein, it is of interest to translate the sequence and to analyze the properties of the putative translation product. Comparison of the peptide sequence so obtained against databases of known proteins and searches for protein motifs could shed light on the function of the gene that was originally sequenced. If further sequencing or cloning is to be carried out, it might be necessary to design primers for sequencing or PCR, or to choose probes for library screening. All this and more can be done with the help of MacVector. The newest version, MacVector 6.5, has just been released, and like earlier versions, it comes with a separate contig assembly module called AssemblyLIGN. This module is useful for assembling short sequences obtained from various regions of a larger sequence into a single contiguous stretch, or contig. Thus, if the sequence of a 2-kilobase complementary DNA was obtained in six sequencing runs of overlapping 400–base pair segments, the results could be imported into AssemblyLIGN, which would find the overlaps and assemble them into a single sequence. Both programs install from a single CD-ROM in minutes, and take about 19 megabytes of hard-disk space. Users without a CD-ROM drive may install the program over a network from another computer. MacVector has a copy protection device; however, one may purchase a site license, monitored by a KeyServer, for a specified number of users.

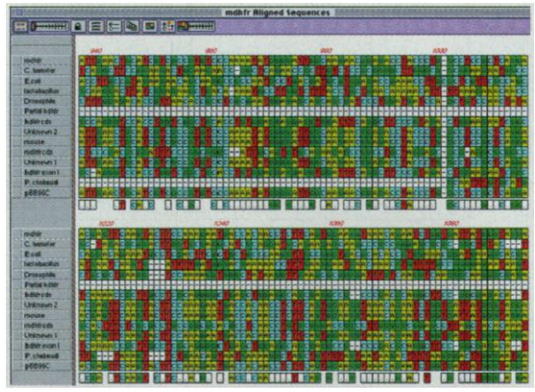

Fig. 1. Results of CLUSTAL W alignments can be formatted with the Multiple Sequence Alignment Editor.

The program is accompanied by a slim, 32-page booklet to orient new users and by a larger user guide filled with helpful tips and detailed information on the various analyses. The program itself contains a Tutorial folder that leads the novice through different types of analyses, step by step. In the best traditions of Macintosh software, the program is easy to use and it is a simple matter for anyone used to general Macintosh conventions to quickly delve into actual sequence analysis.

The author is in the Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA. E-mail: rajagopi@ucs.orst.edu

Entering Data

The first step, of course, is entering the sequence to be analyzed into the program. The data may be entered manually into the sequence window, either by typing or via a gel reader, and saved as a MacVector file. Two MacVector features are helpful in this process. First, keys on the numeric pad of the keyboard can be assigned nucleotide codes, so that sequence data can be entered using one hand (a small Help window will remind you of the IUPAC codes, if you have forgotten). Second, a proofreader option allows the sequence to be read back to the user as it is being entered, as well as after entry is completed. This proofreader, which comes with a choice of a male or a female voice to read the sequence back to the user, is invaluable. If the sequence is already available as a TEXT file, in one of several formats, it can be directly opened by MacVector. Among the compatible formats listed are GenBank flat file, IG_SUITE, CODATA, EMBL, PearsonFASTA and FASTP, DNA Strider ASCII, Staden, and GCG. In addition, MacVector can open files exported in MacVector format from OMIGA, which is Oxford Molecular's sequence analysis program for Windows. Once sequences are in MacVector they may be saved as TEXT files in non-MacVector formats, including most of those listed above. The sequence window has numerous handy little controls represented by a row of icons along the top. These include switches that permit viewing of the sequence as DNA or RNA, circular or linear, and single or double-stranded, as well as control of the features table, the graphic feature map, and sequence read-back (voice verifying). As in any word processor, residues may be added, deleted, or inserted, while whole blocks of sequence may be cut, copied, and pasted between files. Sequences that are saved as MacVector files may be imported into AssemblyLIGN for assembly into a contig, if needed. This is very simply accomplished. Once a contig is assembled, this sequence may be saved for further analysis. Sequences may be reversed, converted to the complementary strand, or reversed and complemented. Text annotations with information about the sequence may also be added, edited, or deleted. Special features of the sequence may be entered in the Features table.

Search Capabilities

MacVector will search sequence files for sites and motifs of various kinds, such as restriction enzyme recognition sites, proteolytic sites, and special nucleic acid or protein motifs. For restriction sites, lists of en-

zymes available from various sources are provided, and one may choose a subset of these enzymes to search the sequence. The program will also look for restriction sites based on user-defined criteria, such as an upper limit on the number of cuts or production of a 3′ or 5′ overhang or blunt end. The analysis is dazzlingly fast, and display options for the results include lists of cutters, noncutters, restriction maps, annotated sequences, and fragment size predictions. The map for displaying the results may be circular or linear and can be edited extensively to produce figures of excellent quality. MacVector also provides lists of common nucleic acid and protein motifs, such as Shine-Dalgarno sequences, binding sites for various transcription factors, nuclear localization signals, phosphorylation sites, leucine zippers, and the like. Each list is editable; that is, one may add new motifs or modify existing ones. Sequences may be searched for one or more of these motifs, and the identified motifs may be displayed as a list or on a map of the sequence.

A nucleic acid sequence may also be screened for PCR primers, sequencing primers, or hybridization probes. The program will present the results of its search, together with information on length and percent guanosine-cytosine (G-C) content of each primer, melting temperature (T_m), primer pair T_m difference, primer-dimer formation, hairpins, product size, and so on. If the user has already chosen primers, these sequences may be analyzed for the characteristics listed above in seconds. This analysis can help avoid the trouble, time, and expense of making primers that will not perform optimally. MacVector will also assist in the design of a hybridization probe from a protein sequence. If the user wishes to clone a gene by using information from the amino acid sequence of the gene product, the program can reverse-translate the protein sequence, then scan the DNA sequence obtained to identify regions of least codon degeneracy to be used as probes. It will provide the probe sequences, G-C content, and T_m, as well. Anyone who has performed this tedious task manually will appreciate the convenience of this feature.

Nucleic acid sequences that are to be searched for protein coding regions may be analyzed by using the open reading frame (ORF) finder. The user has the choice of designating start and stop codons or, for sequences that are at least 200 nucleotides long, of using a statistical method that employs Fickett's TESTCODE algorithm. MacVector will also create codon bias tables by searching a database for sequences from a given organism, locating known coding regions, and calculating the codon usage in those coding regions. This feature is especially useful for minimizing degeneracy when designing probes or PCR primers based on protein sequence data. Once an ORF has been identified, it may then be translated by using either the default genetic code or a genetic code of the user's choice. The peptide sequence obtained from this step may then be examined for properties such as secondary structure, regions of hydrophilicity/hydrophobicity, antigenicity, flexibility, and so on. These analyses are found under the Protein Analysis Toolbox, together with options for calculating amino acid composition, isoelectric point, and molecular weight. In a chapter that is exemplary in its clarity, the user guide provides an introduction to each analysis and the algorithms used. A reference list cites the original papers describing these methods.

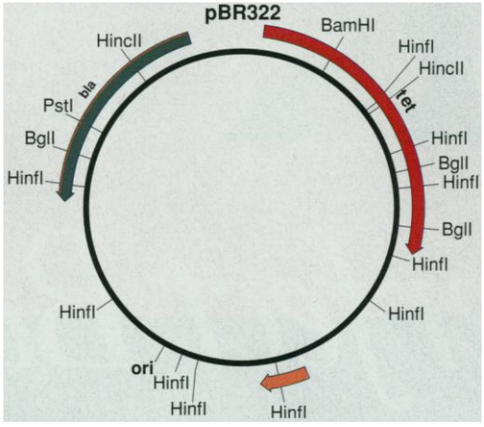

Fig. 2. High-quality maps and graphics are just one feature of the latest version of MacVector.

Equally useful and well-written is the section of the manual that deals with nucleic acid and protein sequence comparison and alignment. At some point, most molecular biologists find themselves in the position of needing to carry out sequence similarity ("homology") searches or sequence alignments. MacVector will perform paired sequence comparisons by Pustell matrix analysis and will provide the user with a plot of the results. In the user guide, novices will find simple explanations of terms like hash size, jump parameters, and *k*-tuple, as well as help in customizing searches and interpreting results. Query sequences may also be compared with those in databases available over the Internet on the National Center for Biotechnology Information (NCBI) server. This method uses a heuristic search algorithm, Basic Local Alignment Search Tool (BLAST) to find sequences similar to nucleic acid or protein query sequences. One especially useful feature of these searches is the ability to extract Medline abstracts pertaining to the search results and to save them to the computer's hard disk. Detailed instructions make it easy for the Internet-phobic to carry out these operations with a minimum of trauma. For multiple sequence alignments, MacVector offers the CLUSTAL W algorithm. This method can be used to simultaneously align several nucleic acid or protein sequences. The results may be displayed as pairwise or multiple alignments, or as a dendrogram. The dendrogram that is generated, however, is not a true phylogenetic tree and should not be mistaken for one. Aligned sequences may be output as high-resolution PICT files and may be edited in standard graphics applications. There are a number of choices available for highlighting conserved residues and for using color to show blocks of sequence identity or similarity. With the help of the Multiple Sequence Alignment Editor, results from CLUSTAL W alignments can be presented clearly and attractively (Fig. 1).

Conclusions

In summary, MacVector is an outstanding sequence analysis program that combines extraordinary ease of use with a comprehensive and powerful toolkit. It is an especially good choice for labs with students or novice technicians, because it is so easy to learn that even the uninitiated can soon be productively engaged in sequence analysis. Users of older versions of MacVector will be pleased to see the considerable improvement in the quality of graphics in MacVector 6.5 (Fig. 2). It is possible to generate high-quality maps and figures, although the program did crash several times during construction of a circular plasmid map. Another new feature is the ability to test user-designed primers for PCR and sequencing. The documentation accompanying the program indicates that some of the minor, but irritating, bugs in earlier versions have been fixed. These changes have made an already powerful and user-friendly program even better than before. A bonus for the worriers among us is the assurance that MacVector 6.5 has been checked for Year 2000 compliance. The publisher assures us that "neither performance nor functionality will be affected by dates before, during, or after the year 2000."

Software Availability

Oxford Molecular Group, Inc., 2105 South Bascom Avenue, Suite 200, Campbell, CA 95008; telephone number: 408-879-6300; fax: 408-879-6302; toll-free: 800-876-9994; e-mail: products@oxmol.com; and url: www.oxmol.com. Price is $2950; $395 or $795 for upgrade.