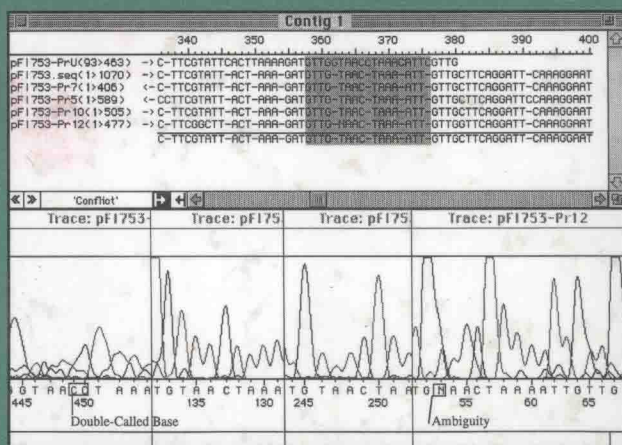


Methods in Molecular Biology™

Volume 70

SEQUENCE DATA ANALYSIS GUIDEBOOK

Edited by
Simon R. Swindell



Humana Press

METHODS IN MOLECULAR BIOLOGY™

Sequence Data Analysis Guidebook

Edited by

Simon R. Swindell


Nottingham University, Nottingham, UK

Humana Press  Totowa, New Jersey

© 1997 Humana Press Inc.
999 Riverview Drive, Suite 208
Totowa, New Jersey 07512

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise without written permission from the Publisher. Methods in Molecular Biology™ is a trademark of The Humana Press Inc.

All authored papers, comments, opinions, conclusions, or recommendations are those of the author(s), and do not necessarily reflect the views of the publisher.

This publication is printed on acid-free paper. 
ANSI Z39.48-1984 (American Standards Institute)
Permanence of Paper for Printed Library Materials.

Cover illustration: Fig. 12 in Chapter 6, "SEQMAN: *Contig Assembly*," by Simon R. Swindell and Thomas N. Plasterer.

Cover design by Patricia F. Cleary.

For additional copies, pricing for bulk purchases, and/or information about other Humana titles, contact Humana at the above address or at any of the following numbers: Tel.: 201-256-1699; Fax: 201-256-8341; E-mail: humana@interramp.com

Photocopy Authorization Policy:

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Humana Press Inc., provided that the base fee of US \$5.00 per copy, plus US \$00.25 per page, is paid directly to the Copyright Clearance Center at 222 Rosewood Drive, Danvers, MA 01923. For those organizations that have been granted a photocopy license from the CCC, a separate system of payment has been arranged and is acceptable to Humana Press Inc. The fee code for users of the Transactional Reporting Service is: [0-89603-358-9/97 \$5.00 + \$00.25].

Printed in the United States of America. 10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging in Publication Data

Main entry under title:

Methods in molecular biology™.

Sequence data analysis guidebook/edited by Simon R. Swindell.

p. cm.—(Methods in molecular biology™; 70)

Includes index.

ISBN 0-89603-358-9 (alk. paper)

1. Nucleotide sequence—Data processing. I. Swindell, Simon R. II. Series: Methods in molecular biology™ (Totowa, NJ); 70.

[DNLM 1. Sequence Analysis—methods. 2. Software.W1 ME9616J v. 70 1997/QH 441 S479 1997]

QP624.S47 1997

574.87'328'0285—dc20

DNLM/DLC

for Library of Congress

96-44660
CIP

Preface

Computers have revolutionized the analysis of sequencing data. It is unlikely that any sequencing projects have been performed in the last few years without the aid of computers. Recently their role has taken a further major step forward. Computers have become smaller and more powerful and the software has become simpler to use as it has grown in sophistication. This book reflects that change since the majority of packages described here are designed to be used on desktop computers.

Computer software is now available that can run gels, collect data, and assess its accuracy. It can assemble, align, or compare multiple fragments, perform restriction analyses, identify coding regions and specific motifs, and even design the primers needed to extend the sequencing. Much of this software may now be used on relatively inexpensive computers. It is now possible to progress from isolated DNA to database submission without writing a single base down.

To reflect this progression, the chapters in our *Sequence Data Analysis Guidebook* are arranged, not by software package, but by function. The early chapters deal with examining the data produced by modern automated sequencers, assessing its quality, and removing extraneous data. The following chapters describe the process of aligning multiple sequences in order to assemble overlapping fragments into sequence contigs to compare similar sequences from different sources. Subsequent chapters describe procedures for comparing the newly derived sequence to the massive amounts of information in the sequence databases.

The acquisition of sequence data is now only the beginning of the process of analysis. Therefore, the later chapters of *Sequence Data Analysis Guidebook* are dedicated to software that allows the data to be examined for specific "features." This includes performing restriction analysis, searching for open reading frames, calculating the translation products of open reading frames, and performing detailed analyses of the "proteins." The software described is sophisticated enough to produce detailed graphic representations of the data showing features such as restriction sites, coding regions, and promoters. It is also able to create simulations of electrophoretic patterns produced by digestion of DNA or protein.

The design of primers either for continued sequencing or manipulation of the analyzed sequence by the polymerase chain reaction is an important part of the analysis process. Several programs designed to aid primer design are described.

Once the sequence has been analyzed as fully as possible, the data is likely to be published. To aid in this, several of the packages described allow the production of complex, publication-quality figures. Since many journals now insist that published sequences be submitted to one of the major databanks, the final chapter deals with the various ways now available to achieve this.

Sequence Data Analysis Guidebook assembles examples of how optimally to use some of the most popular software available today. The packages dealt with have been selected because they are in common use in many sequencing laboratories around the world, and should, therefore, represent a useful reference source for many researchers.

Each of the authors has used these packages extensively: The Notes section found at the end of each chapter is a key feature of this book and contains information derived from their invaluable personal experience.

I would like to acknowledge the efforts of all the contributing authors.

Simon R. Swindell

Contributors

- CATHERINE ARNOLD • *Virus Reference Division, Central Public Health Laboratory, London, UK*
- JONATHAN P. CLEWLEY • *Virus Reference Division, Central Public Health Laboratory, London, UK*
- JONATHAN A. EISEN • *Department of Biological Sciences, Stanford University, Stanford, CA*
- JUAN JOSE ESTRUCH • *CIBA Agricultural Biotechnology, Research Triangle Park, NC*
- TOMAS P. FLORES • *EMBL Outstation—The EBI, Cambridge, UK*
- TRACY L. HAGEMANN • *Department of Immunology, Rush Medical School, Chicago, IL*
- ROBERT A. HARPER • *EMBL Outstation—The EBI, Cambridge, UK*
- SAU-PING KWAN • *Department of Immunology, Rush Medical School, Chicago, IL*
- STEVEN R. PARKER • *Applied Biosystems Division, Perkin Elmer, Foster City, CA*
- THOMAS N. PLASTERER • *DNASTAR, Inc., Brighton, MA*
- BENNY SHOMER • *EMBL Outstation—The EBI, Cambridge, UK*
- EUGENE G. SHPAER • *Applied Biosystems Division, Perkin Elmer, Foster City, CA*
- SIMON R. SWINDELL • *Department of Biochemistry, Queen's Medical Center, Nottingham University, Nottingham, UK*
- PHIL TAYLOR • *MRC Reproductive Biology Unit, Center for Reproductive Biology, Edinburgh, UK*
- BRUCE R. TROEN • *MSRBII, Ann Arbor, MI*

Contents

Preface v

Contributors xi

1 GeneJockeyII: *Entering and Editing Sequences*,
Phil Taylor 1

2 The Genetic Data Environment: *A User Modifiable and Expandable
Multiple Sequence Analysis Package*,
Jonathan A. Eisen 13

3 ABI Analysis: *Manipulation of Sequence Data
from the ABI DNA Sequencer*,
Tracy L. Hagemann and Sau-Ping Kwan 39

4 SeqEd: *Manipulation of Sequence Data and Chromatograms
from the ABI DNA Sequencer Analysis Files*,
Tracy L. Hagemann and Sau-Ping Kwan 55

5 From ABI Sequence Data to LASERGENE'S EDITSEQ,
Catherine Arnold and Jonathan P. Clewley 65

6 SEQMAN: *Contig Assembly*,
Simon R. Swindell and Thomas N. Plasterer 75

7 GeneJockeyII: *DNA Sequencing and Fragment Assembly*,
Phil Taylor 91

8 AutoAssembler Sequence Assembly Software,
Steven R. Parker 107

9 MEGALIGN: *The Multiple Alignment Module of LASERGENE*,
Jonathan P. Clewley and Catherine Arnold 119

10 GeneJockeyII: *Pairwise Sequence Comparison*,
Phil Taylor 131

11 GeneJockeyII: *Multiple Alignment of Homologous Sequences*,
Phil Taylor 137

12 Sequence Navigator: *Multiple Sequence Alignment Software*,
Steven R. Parker 145

13 The European Bioinformatics Institute: *Network Services*,
Tomas P. Flores and Robert A. Harper 155

14	Gene Assist: <i>Smith-Waterman and Other Database Similarity Searches and Identification of Motifs</i> , Eugene G. Shpaer	173
15	GENEMAN of LASERGENE, Jonathan P. Clewley	189
16	GeneJockeyII: <i>Database Searching</i> , Phil Taylor	197
17	GeneJockeyII: <i>Restriction Analysis</i> , Phil Taylor	213
18	GeneJockeyII: <i>Translation and Open Reading Frame Analysis</i> , Phil Taylor	221
19	PROTEAN: <i>Protein Sequence Analysis and Prediction</i> , Thomas N. Plasterer	227
20	MAPDRAW: <i>Restriction Mapping and Analysis</i> , Thomas N. Plasterer	241
21	The Gene Construction Kit: <i>DNA Sequence Analysis and Presentation</i> , Bruce R. Troen	257
22	GeneJockeyII: <i>Primer Design</i> , Phil Taylor	273
23	OLIGO: <i>Primer Selection</i> , Juan Jose Estruch	279
24	PRIME: <i>Primer Selection</i> , Juan Jose Estruch	287
25	PRIMERSELECT: <i>Primer and Probe Design</i> , Thomas N. Plasterer	291
26	The European Bioinformatics Institute: <i>Submission and Updating of Sequence Databases</i> , Tomas P. Flores and Benny Shomer	303
	Index	319

GeneJockeyII

Entering and Editing Sequences

Phil Taylor

1. Introduction

Entering sequence by hand is a tedious and error-prone process. In general, if the sequence that you need is available in any electronic form, you should be able to import it into GeneJockeyII without having to retype the data. For example, most sequences published in research papers are normally accompanied by a GenBank/EMBL accession number, which allows you to retrieve the sequence from the GenBank CD-ROM or from a remote networked database. If, however, you have no option but to type the required sequence (for example, if you are reading sequence by hand from a manual sequencing gel), GeneJockeyII provides powerful facilities to do so, and to check the accuracy of the entered data. Sequence data in GeneJockeyII is simple text, displayed in capitals, and behaves just as text does in any word processor. All the standard editing commands act in the way in which you expect them to act, and you may use fonts, styles, and colors to draw attention to parts of your sequence, just as you would when editing ordinary text.

2. Materials

1. Hardware: GeneJockeyII requires a Macintosh with ColorQuickdraw in ROM (this excludes the Macintosh plus [and older machines], the SE, the PowerBook 100, and the Macintosh Portable). The program also requires system 7.0 or later, and at least 2 Mb of available memory. A color display capable of showing 256 colors is helpful but not essential.
2. Software: For the operations described in this chapter, you need only the GeneJockeyII program itself. For operations described in later chapters, you will need some additional files supplied with the program. You would normally install

From: *Methods in Molecular Biology*, Vol. 70: *Sequence Data Analysis Guidebook*
Edited by: S. R. Swindell Humana Press Inc., Totowa, NJ

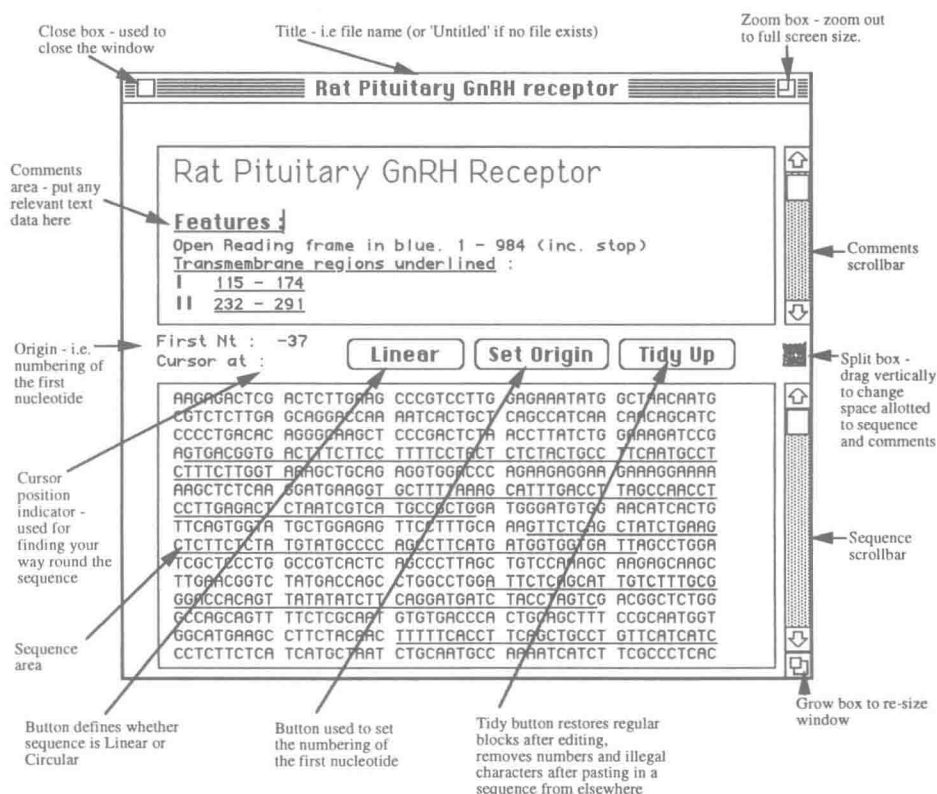


Fig. 1. Anatomy of a GeneJockeyII sequence window.

GeneJockeyII on your hard drive by simply copying all the files supplied into a single folder. When running on a Power Macintosh, the GeneJockey Helper file should be present in the same folder. The native-code resources in this file run about ten times faster than the code in the main program, and since multiple alignment is a time-consuming process, the extra speed is very helpful. GeneJockeyII is licensed for use only on a single-user basis, but is not copy-protected.

3. Methods

3.1. Sequence Entry

1. Start up the program by double-clicking on the GeneJockeyII icon. The program offers three kinds of windows in which you may enter and edit text. For this reason, the New command in the File menu is hierarchical, offering you the choice of a new nucleotide sequence window, peptide sequence window, or a plain text window. We will start by opening a nucleotide sequence window and entering a DNA sequence (*see* Note 1). Fig. 1 shows a nucleotide sequence window.

2. Use the **New > Nucleotide** sequence command to open the window. Note that the window title is **Untitled 1**. As is usual with Macintosh programs, the window will not be given a title until you save it to disk (see Note 1).
3. Use the **Save as...** command from the **File** menu to save your new window before you start typing.
4. Give the file a suitable name for the sequence you are going to enter.
5. When the file is saved, click on the empty sequence box to place the insertion point at the top left of the box.
6. Start typing your DNA sequence. Note that the program converts text that you type into this box to uppercase (see Notes 2,3).
7. Next, select **Speak on Entry** from the **Edit** menu. Continue typing. Each time you hit a key, the machine will speak the corresponding letter. This is very helpful if you are not a touch typist, because it means that you do not have to look at the screen to check what you type. You can turn this facility off again using the same command.
8. Select **Tidy up** to format the sequence into blocks of 10 nucleotides (see Note 4).
9. Once you have typed in a few lines of sequence use the **Save** command to update the disk file. It is always a good idea to save sequences frequently when typing, in case of accidents. You should make sure your sequence is saved before carrying out the operations in the next paragraph.
10. Type a few more bases and look at the **Revert to Original** and **Undo** commands (see Notes 5,6).

3.2. Switching Between Circular and Linear Sequences

1. Of the three buttons at the center of the screen, the left-hand button currently reads **Linear**. When you click on it, the legend changes to **Circular**. The button toggles between these two states, and the legend indicates the current conformation of the sequence. The difference between linear and circular sequences is for the most part trivial, affecting only the restriction enzyme analysis, in which it is important to deal correctly with restriction sites that span the origin of circular sequences (i.e., where part of the site is at the top left of the display at position 1, and the other part at the very end).
2. Click on the button again to return the sequence to the linear state.

3.3. Changing the Origin Point

1. Click on the **Set Origin** button. You will see a dialog box that asks you for the number of the first nucleotide in the sequence and tells you that you may enter any number between 32 and -32 K, except zero (see Note 7).
2. Enter a small negative number, such as -20, and click on **OK**. The **First Nt:** legend at center left now reads -20 to remind you of the current numbering, and if you run the cursor along the top line of the sequence, you will see that the numbering jumps from -1 to +1 without using zero.
3. Click on the **Set Origin** button again and set the origin back to 1.
4. Now, make the sequence circular, and if you have made any changes, save the sequence again. (If you can not remember whether you have made any significant

changes, pull down the **File** menu and look at the **Save** command. If it is disabled then you do not need to save.)

5. Now use the **Set Origin** button again. The effect of changing the origin of a circular sequence is quite different, since by convention the origin of a circular sequence is always shown at the top left of the display. If you set the origin to -20 , the sequence will be rotated so that the last 20 nucleotides are brought to the beginning, with the nucleotide that was twentieth from the end of the sequence displayed at the top left, and numbered 1. Remember that there is no **Undo** command for this, so it is a good idea to make sure the sequence is saved in case you make a mistake with the numbering. You can then use the **Revert** command to restore the original display. The effect of circularizing a linear sequence whose origin is not the first nucleotide displayed is similar, and the same caution applies here.

3.4. Verifying the Sequence

Entry of sequences at the keyboard is an error-prone process, and if you wish to be certain that the sequence you have entered is correct it is necessary to use some form of verification. GeneJockeyII offers you two methods of verifying sequences: **Verify by Speaking** and **Verify by Typing**. Both commands are found in the **Edit** menu.

1. First, click at the top left of the sequence to set the insertion point at the beginning (or just before the part of the sequence you wish to check).
2. Select **Verify by Speaking**. The computer will speak the first 10 bases of the sequence, permitting you to check that you have entered them correctly. Hit the space bar or any other printing key to start reading the next 10 bases. If you wish to move quickly around the sequence, use the left or right arrow keys to move forward or back 10 bases, or the up or down arrow keys to move one line up or down (see Note 8).
3. Set the insertion point back to the beginning of the section you wish to check.
4. Select **Verify by Typing** (see Note 8).
5. Start retyping the sequence. As you type each base, the selection moves one place forward. If you type a base that does not match the sequence you entered originally, the machine will beep and the selection will not move on.
6. In order to correct the error, type Command-period and the machine will return control to you with the incorrect base already selected for changing.
7. Type the correct base then reissue the **Verify by Typing** command to continue verification (since this is a keyboard-orientated operation you will find it quicker to use the Command-T equivalent to restart verification). As before, you may use the arrow keys to move around rapidly during verification, and the machine will exit from the mode automatically when you reach the end of the sequence.

3.5. Annotating Sequences

You can insert notes and comments on your sequence in the upper text box of the window. Only one of the text boxes is active at any time, indicated by the flashing insertion point.

12. Set the insertion point on the cut site, i.e., between the G (at 701) and the following A.
13. Click on the test sequence window to bring it back to the front.
14. Select the whole sequence by means of the **Select All** command from the **Edit** menu. (You could also do this by dragging across the whole sequence, or by setting the insertion point at the beginning and shift-clicking at the end.) So that we will be able to identify the insert when we have made the construct, it is a good idea to label it now.
15. Use the **Color...** command from the **Text** menu to put the sequence into a contrasting color (see Note 10).
16. Next, copy the entire sequence onto the clipboard by means of the **Copy** command from the **Edit** menu.
17. Bring the vector sequence window back to the front. If you do this by clicking on it, be careful to click only once, or you may shift the insertion point from the place where you left it. Check that it is still after the G at 701.
18. Paste the test sequence in using the **Paste** command from the **Edit** menu.
19. Click on the **Tidy up** button to reformat the sequence.
20. **Save** it under a suitable name.

There—you have just ligated a test sequence into a vector—I bet you wish it was that simple in the real world!

3.7. Inverting Sequences

Suppose that we have only the construct sequence to work with, but we decide that the wrong strand of DNA has been inserted into the vector, and we need to take it out, invert it (i.e., generate the opposite strand), and put it back again. First, we have to select the insert, which is now in the middle of the pBluescript sequence. We know where the beginning is, just after the *EcoRI* site at 701, so we only need to locate the end. We could find that numerically by adding the length of the test sequence to 701, or we could simply scroll down the screen to see where the color changes, but we will search again for the second *EcoRI* site, which now marks the end of the insert.

1. Set the insertion point at the beginning of the sequence
2. Select the **Find Same** command. This simply repeats the previous search, finding the original site.
3. Repeat the **Find Same** command to find the second *EcoRI* site.
4. Set the insertion point just before the G of the second site.
5. Scroll back to the first site at 701. Hold down the shift key while you click after the C of the first site. The whole of the insert will then be selected. (In GeneJockeyII, the cursor display remains active while you drag, so you could also just drag across the part of the sequence that you want, watching the numbers to see when you get to the right place. Yet another alternative would be to use the **Select...** command from the **Find** menu and specify numerically the region of sequence you want selected.)

12. Set the insertion point on the cut site, i.e., between the G (at 701) and the following A.
13. Click on the test sequence window to bring it back to the front.
14. Select the whole sequence by means of the **Select All** command from the **Edit** menu. (You could also do this by dragging across the whole sequence, or by setting the insertion point at the beginning and shift-clicking at the end.) So that we will be able to identify the insert when we have made the construct, it is a good idea to label it now.
15. Use the **Color...** command from the **Text** menu to put the sequence into a contrasting color (*see* Note 10).
16. Next, copy the entire sequence onto the clipboard by means of the **Copy** command from the **Edit** menu.
17. Bring the vector sequence window back to the front. If you do this by clicking on it, be careful to click only once, or you may shift the insertion point from the place where you left it. Check that it is still after the G at 701.
18. Paste the test sequence in using the **Paste** command from the **Edit** menu.
19. Click on the **Tidy up** button to reformat the sequence.
20. **Save** it under a suitable name.

There—you have just ligated a test sequence into a vector—I bet you wish it was that simple in the real world!

3.7. Inverting Sequences

Suppose that we have only the construct sequence to work with, but we decide that the wrong strand of DNA has been inserted into the vector, and we need to take it out, invert it (i.e., generate the opposite strand), and put it back again. First, we have to select the insert, which is now in the middle of the pBluescript sequence. We know where the beginning is, just after the *Eco*RI site at 701, so we only need to locate the end. We could find that numerically by adding the length of the test sequence to 701, or we could simply scroll down the screen to see where the color changes, but we will search again for the second *Eco*RI site, which now marks the end of the insert.

1. Set the insertion point at the beginning of the sequence
2. Select the **Find Same** command. This simply repeats the previous search, finding the original site.
3. Repeat the **Find Same** command to find the second *Eco*RI site.
4. Set the insertion point just before the G of the second site.
5. Scroll back to the first site at 701. Hold down the shift key while you click after the C of the first site. The whole of the insert will then be selected. (In GeneJockeyII, the cursor display remains active while you drag, so you could also just drag across the part of the sequence that you want, watching the numbers to see when you get to the right place. Yet another alternative would be to use the **Select...** command from the **Find** menu and specify numerically the region of sequence you want selected.)

6. **Copy** the insert onto the clipboard.
7. Use the **New > Nucleotide Sequence** command to generate a new sequence window.
8. Paste the sequence into it and **Tidy** it.
9. Select **Invert** from the **Modify** menu. The program opens a new window containing the inverted sequence (see Note 11).
10. Use **Select All** to change the color as before, if you wish.
11. **Copy** the entire sequence.
12. Pull the window containing the construct back to the front. Since we now have several windows open, it is easier to do this by means of the **Windows** menu than by trying to find it by moving the windows around on the screen. The part of the sequence that represents our original insert is still selected.
13. **Paste** the inverted sequence, and it will replace the original.
14. **Tidy up** the sequence.

We are now finished with the windows that we currently have open, so close them all. To do this, hold down the Option key while clicking in the close box of the front window. The program will close all the windows in turn, prompting us as it does so to save any new work.

4. Notes

1. Using the **New** command offers three alternatives. One is for creating a new nucleotide sequence. The second is for creating a new peptide sequence. Peptide sequences are entered in precisely the same way as nucleotide sequences, and a peptide sequence window looks just like a nucleotide sequence window, the only obvious difference being that the origin prompt at center left reads "First AA:" rather than "First Nt:." You will notice some differences when you come to use the modification and analysis commands, however, since different menu commands will be enabled depending on what type of window is foremost on the screen.

Peptide sequences are entered in single letter code and represented in uppercase characters only. There are no wildcard characters. The type of window you choose specifies whether the program will treat the sequence as DNA or protein, and there is very little to prevent you from entering the wrong kind of sequence into a window (there is no way for the program to distinguish between a short DNA sequence and the equivalent set of characters representing a peptide consisting entirely of alanine, cysteine, glycine, and threonine, for example), so be careful when using the **New** command to ask for the correct window type for the sequence you intend to enter.

A third type of window that may be obtained with the **New** command is a plain text window. This has a single scroll bar and is 80 characters wide. There is a title area at the top that holds a single line of text and initially reads "New text window." This title string is not directly editable, but may be changed via a dialog box obtained by clicking in this area. The remainder of the window acts as a plain text area, and is useful for general purpose editing. Many of the analyses that GeneJockeyII performs display their results in text windows, and you may edit such results before printing or saving them.

2. GeneJockeyII only handles sequences consisting of uppercase symbols. Note that when you reach nucleotide number 10, and any multiple of 10 thereafter, the program will automatically insert a space or return so that the sequence is displayed in blocks of 10. In a nucleotide sequence window, you may use the symbols A, C, G, and T, plus the standard degenerate symbols that are used to represent the case in which a particular position may be occupied by more than one base. U is not a legal character, so RNA sequences should be entered as DNA. If you type an illegal character you will get a dialog box displaying the complete list of these characters. For example, type in an X to see this. You can also see the display of permitted degenerate codes at any time by selecting the **Show Wildcards...** command from the **Edit** menu. You can dismiss the Wildcards dialog either by clicking on the **Cancel** button or by clicking on any of the buttons that display the degenerate codes; in the latter case the dialog causes that code to be inserted into the sequence at the current selection point.
3. When entering DNA sequences you will make extensive use of the A, C, G, and T keys, and it is most convenient to have these keys close together so that you can enter the data with one hand and not have to look at the keyboard. Use the **Re-Assign Keys...** command from the **Edit** menu to do this. Because I am right-handed, I normally reassign the keys U, I, O, and P to give me A, C, G, and T, respectively. This has the advantage that none of U, I, O, or P are degenerate codes, so I will never want to use them for their original symbols within a DNA sequence, and they are close enough on the keyboard to the delete key that if I make a mistake I can backspace over it without taking my eyes off the gel or sequence from which I am reading. If you wish your keyboard always to work in this way, you should click on the **Set Default** checkbox before clicking on **OK** in the dialog. To return the keyboard to normal you should click on the **Standard Layout** button. The reassigned keyboard only applies to DNA sequences; the keyboard will operate normally when you type ordinary text into the comments area of a sequence window or anywhere else.
4. You have probably noticed by now that if you move the mouse cursor across the sequence box the number of the nucleotide beneath the cursor is continuously displayed at center left. This is very helpful for locating a particular nucleotide by number. The calculation of the number does, however, depend on the sequence being formatted correctly in regular blocks of ten. Some operations destroy this regular format, and the function of the **Tidy up** button is to restore order in these cases. For example, suppose you wished to insert an extra block of sequence in the middle of your existing sequence. Place the insertion point in the middle of the sequence by clicking on it. Now type in a few nucleotides. The resulting disorder would not affect any analyses that you later ran on this sequence, since all the analyses ignore the presence of space and return characters, but it looks untidy and spoils the operation of the cursor position display. Click on the **Tidy Up** button to put the sequence back into regular columns. It would have been possible to make the program tidy the sequence after every keystroke, but it would have slowed the operation of the program to an irritating extent, especially when inserting residues near the beginning of a long sequence.

5. If you now wish to restore your sequence to its original state, select the **Revert to Original** command from the **File** menu. This returns the window to the state it was in when you issued the last **Save** command, checking with you first to see if you really want to discard any changes made since then.
6. Another way to reverse any change you have made is to use the **Undo** command at the top of the **Edit** menu. Pull down the menu and look at this command now. It reads **Undo Typing**, and if you use it, all the typing you have done since you placed the insertion point will be removed. **Undo** always shows you what can be undone. Almost all editing operations can be undone, the only exceptions being the three operations performed with the buttons at the center of the screen. It may read **Cannot Undo**, and be disabled (i.e., it is shown in gray, and does not respond if you try to use it). This is because the file has just been loaded or saved, and you have not yet made any changes: There is nothing to undo.
7. **Set Origin** changes the way in which the sequence is numbered, and has different effects depending on whether the sequence is linear or circular. The origin of a linear sequence is position number 1, which may be anywhere on the screen, or indeed outside the sequence displayed. If your sequence represents a small segment of a larger sequence that is itself numbered from 1, the first nucleotide displayed on the screen will have a number >1. If, on the other hand, you wish to set the origin at some feature in the body of the sequence (for example, at the start codon of a translated region), the first nucleotide will have a negative number. By convention, nucleotide numbering does not use zero, so you may not set the origin to zero. Strictly speaking, when you set the origin of a linear sequence, you do not specify the position of the origin itself, but rather the numbering of the first nucleotide.
8. **Verify by Typing** and **Verify by Speaking** are modal commands, i.e., you can not do anything else at the same time, because the menus, scrollbars, and so on, are all inactive. When the program has talked its way to the end of the sequence it will exit automatically from this mode and return to normal operation. If you wish to exit before the end of the sequence is reached (in order to make corrections) you may do so by holding down the command key and simultaneously typing a period. (This is the standard Macintosh abort command: You can stop most operations in GeneJockeyII this way if you change your mind.)
9. The **Find** command in GeneJockey is similar to that in a word processor, but has some special facilities for use with sequences. Since all sequences in GeneJockey are in uppercase, it does not matter whether you type in the target sequence in capitals or lowercase; the program will convert the characters to capitals before searching. You can include degenerate codes in the target sequence, so AATNG will find AATAG, AATCG, AATGG, or AATTG. Likewise, degenerate codes in the search sequence will be honored, so AATTG will find not only AATTG but NATAG, ANTAG, AANAG, and so on. The **Find** command will also permit you to specify a number of allowable mismatches, so you can find sections that are similar to, but not identical to the target sequence. You can also set the program to find the minimum number of mismatches required to produce a match, by means of the **Find Mismatches** button.