CAD Gene™

CAD Gene
Tutorial
Sci Vision, Inc.
128 Spring Street
Lexington, MA 02173

CAD Gene™

Sequence Analysis Program
Version 1.07
for the Apple® Macintosh™

Sci Vision, Inc.

Use of the Software contained in this package is has been provided under a Software License Agreement. Please read it carefully. Unauthorized use of the software and related materials can result in civil damages and criminal penalties.

LIMITED WARRANTY AND DISCLAIMERS

The software and related materials in this package are licensed with a limited 90-day limited Warranty:

A the media in which Software is embodied will not be defective and the programs will be properly recorded upon such media; and
B. the Software will function substantially as described in the Documentation.

If it is determined that the Software does not comply with the above warranties, Licenser's only responsibility and Customer's sole and exclusive remedy for any such breach of warranty or any negligence relating to the Software, including, without limitation, as to any installation, maintenance or enhancements, will be for Licenser to use reasonable efforts to correct technical errors or defects in the Software, and/or at Licenser's sole option to repair or replace all or any portion of the Software. This limited warranty does not cover damage to software resulting from accident, misuse or modification of the Software and is valid only if Licenser receives written notice of breach of warranty within such 90-day period.

NO OTHER REPRESENTATION OR WARRANTIES, WHETHER EXPRESS OR IMPLIED, INCLUDING, WITHOUT LIMITATION, ANY EXPRESS OR IMPLIED WARRANTIES OF GOOD TITLE, NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE ARE MADE BY LICENSER WITH RESPECT TO ANY SOFTWARE, DOCUMENTATION OR SERVICES FURNISHED HEREUNDER. IN NO EVENT WILL LICENSER BE LIABLE TO CUSTOMER OR ANY OTHER PARTY FOR ANY DAMAGES OR LOST PROFITS OR FOR ANY CONSEQUENTIAL, SPECIAL, INCIDENTAL OR PUNITIVE DAMAGES WHICH MAY ARISE IN CONNECTION WITH THIS AGREEMENT OR ANY SOFTWARE, DOCUMENTATION OR SERVICES, INCLUDING BY REASON OF ANY NEGLIGENCE. IN NO EVENT WHATSOEVER SHALL LICENSER HAVE ANY LIABILITY HEREUNDER FOR A CUMULATIVE AGGREGATE AMOUNT IN EXCESS OF LICENSING FEES PAID BY CUSTOMER.

TABLE OF CONTENTS

CAD Gene™ basics

CAD Gene™ is a new direction in the development of interactive software for molecular biology. The software is designed to fully exploit the interactive nature of current desktop computer hardware and to facilitate data analysis and project planning. Its intuitive interface has been designed by working molecular biologists to closely parallel the thinking patterns used by most researchers.

This book is an introduction to using CAD Gene™, and will quickly guide you through a series of simple examples to illustrate how the program is utilized in molecular biology research. If you follow the instructions in this book, you will learn how to

- install and start using CAD Gene™
- open CAD Gene files
- perform restriction digests
- extract restriction fragments from restriction digests
- ligate sequences

What you need

To use CAD Gene™ on your Macintosh, you must have the following equipment and supplies. We strongly recommend a hard disk drive.

- A Macintosh Plus or later computer with a minimum of 1 megabyte of memory. CAD Gene™ is compatible with all current models of Macintosh computers, including the Power Macintosh.
- System 4.1 or later, Finder 5.3 or later. CAD Gene™ is compatible with MultiFinder and System 7.
- The CAD Gene™ master disk.
- A blank, initialized disk with which to make a working copy of the program (not needed if using a hard disk). CAD Gene™ and its associated files take up less than 600 kilobytes of disk space.
- A printer, if hard copy is desired. CAD Gene™ has been tested and is compatible with the LaserWriter, StyleWriter, ImageWriter, and LaserJet printers.

What you need to know

Before you use CAD Gene™ or this book, you should be familiar with your Macintosh and its operation. You should be able to do the following:

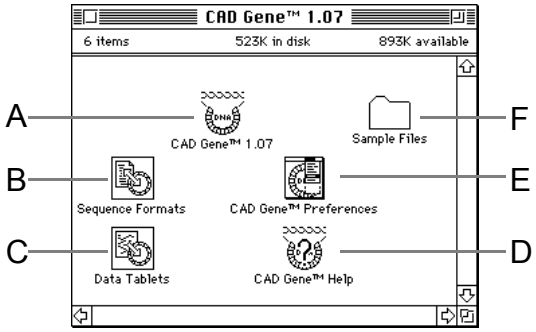
- Turn your computer and any attached equipment on and off
- Use the mouse to select and drag items
- use pull-down menus
- Open and close windows
- Use scroll bars
- Start and quit applications
- Create and use folders
- Copy and move files and folders
- Use the trash can
- Use and copy disks

If you are unfamiliar with any of the above terms or concepts, you should review the *Macintosh Basics* disk and booklet that came with your Macintosh computer.

Additionally, you should be familiar with basic concepts in molecular biology.

Files on the CAD Gene™ master disk

The CAD Gene™ master disk contains the following files and folders.



A. Main program. Referred to as "CAD Gene™" throughout this book. The name of this program on the disk may reflect the current shipping version number of the program. Double click on this file to start CAD Gene™.

B. Sequence Formats file. Contains the information to support foreign (DNA) sequence files, restriction enzyme lists.

C. Data Tablets file. Contains the subroutines to support sequence entry with graphics tablets. This file is not necessary if a data tablet (graphics tablet) will not be used for sequence entry.

D. CAD Gene™ Help file. Contains information for on-line program help.

E. CAD Gene™ Preferences. Contains default settings for restriction digest group, genetic codes, etc. A preferences file must be present for the program to run.

F. Sample files folder. Contains the sequence files described in the CAD Gene™ tutorial.

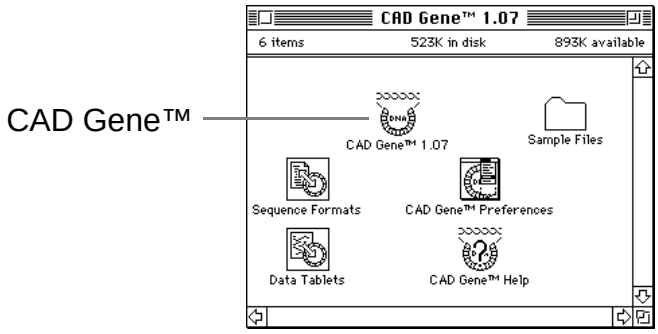
Why so many files?

CAD Gene™ will read sequences that have been saved as TEXT in several different formats -- GenBank, UWGCG, Staden. CAD Gene™ not only reads in the nucleotide sequence from these files, but also reads in all the site information (start sites, coding sequence, splice sites, etc.). CAD Gene™ uses a proprietary internal format in order to be able to process this information more rapidly.

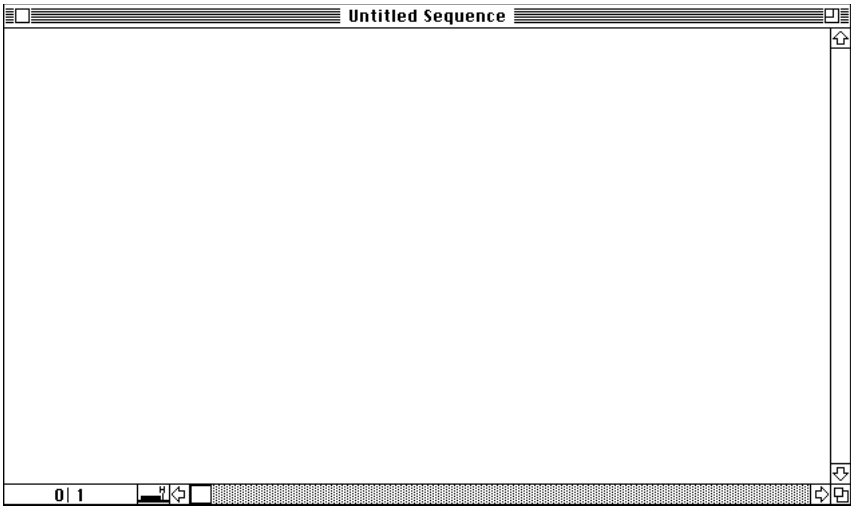
The sequence format specifications (such as GenBank) are subject to frequent changes. Since the information necessary to convert foreign sequence formats into the CAD Gene™ internal format is stored in a separate file. This information can be changed without changing the main program. Genetic Technology Corp. and Sci Vision can update sequence format conversion options independently of updates to the main program, to keep abreast of the latest changes in sequence formats.

Getting to work

In the next several pages, you will be led through a series of examples that illustrate the use of your CAD Gene™ software. The files used in the examples are included in the "Sample Files" folder. The exercises are best completed by following along on your Macintosh. To begin, open the CAD Gene™ program by double-clicking on the CAD Gene™ program icon.



When the CAD Gene™ program is opened directly by double clicking on the program icon, an empty document is opened ready for entering DNA sequence. CAD Gene™ is designed to allow only the entry of valid sequence characters.



Subcloning

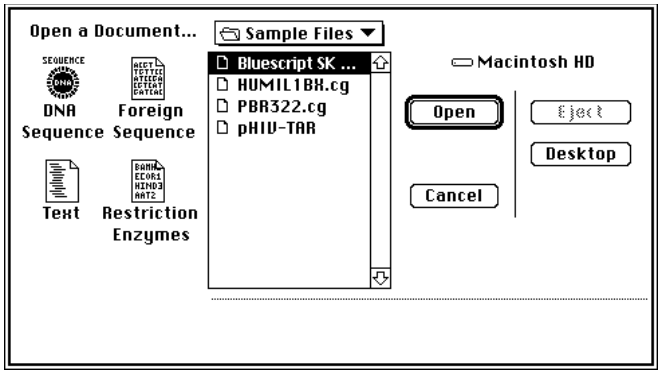
In order to demonstrate how CAD Gene™ is used in planning cloning projects, we will step through a simple sample exercise. In this exercise, we will subclone exon 1 of the human interleukin 1 gene into the commercial cloning vector Bluescript SK (-). The process using CAD Gene™ parallels the procedures that are performed at the lab bench -- we will "extract" a fragment from a restriction digest of human IL1, "extract" a vector fragment from a restriction digest of Bluescript SK(-), and then "ligate" these fragments to produce a new clone.

To begin, we must open the sequence files containing the source files. Select "Open" from the file menu.

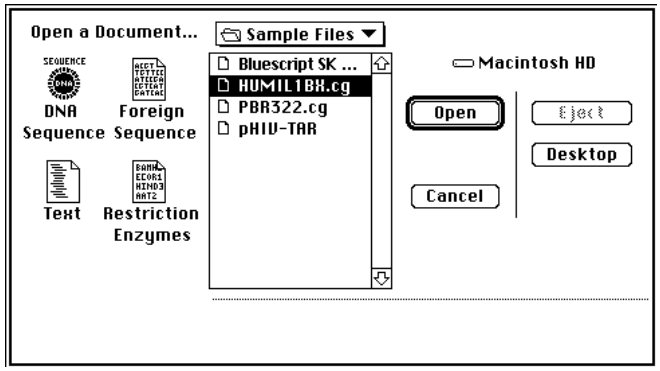
The "Open" dialog will appear. Open the "Sample Files" folder located in the CAD Gene™ folder. The files that appear when the "Sequence" button is selected are valid CAD Gene™ format sequence files. CAD Gene™ supports a proprietary file format. The proprietary file format allows

for the content of sequence files to be controlled so that they contain only valid sequence characters. The sequence in CAD Gene™ files are also encoded in such a way as to facilitate their rapid analysis.

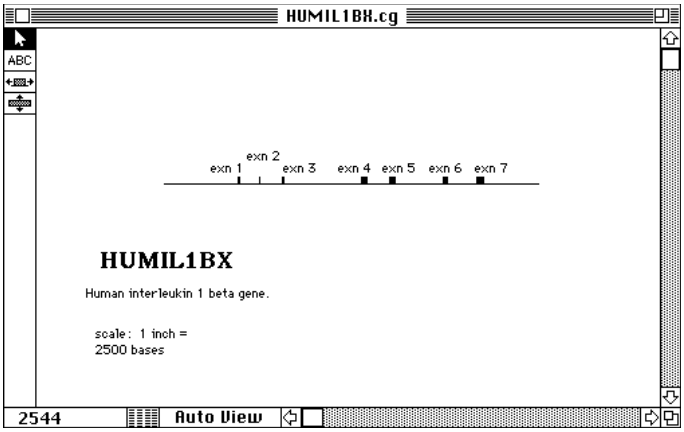
CAD Gene™ also supports inputting sequences saved as TEXT files from other sources ("Foreign Sequences"), plain text files ("Text") and restriction enzyme lists saved as TEXT files ("Restriction Enzymes").



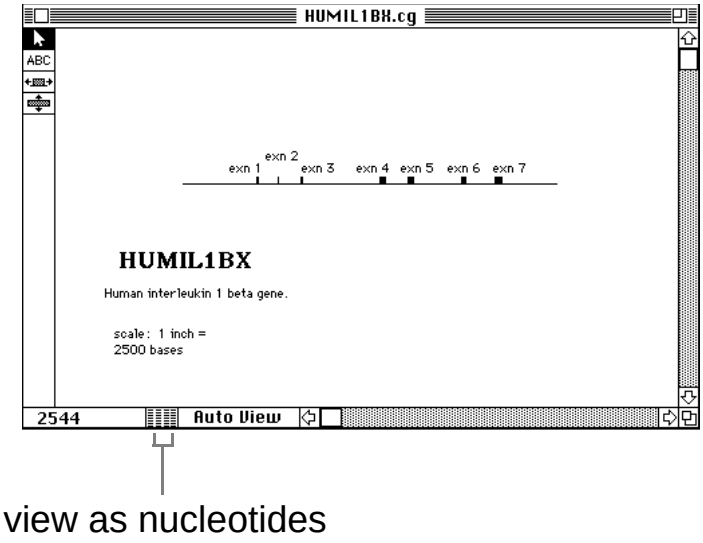
Select "HUMIL1BX.cg" and then select the "Open" button.



The sequence "HUMIL1BX" will appear.

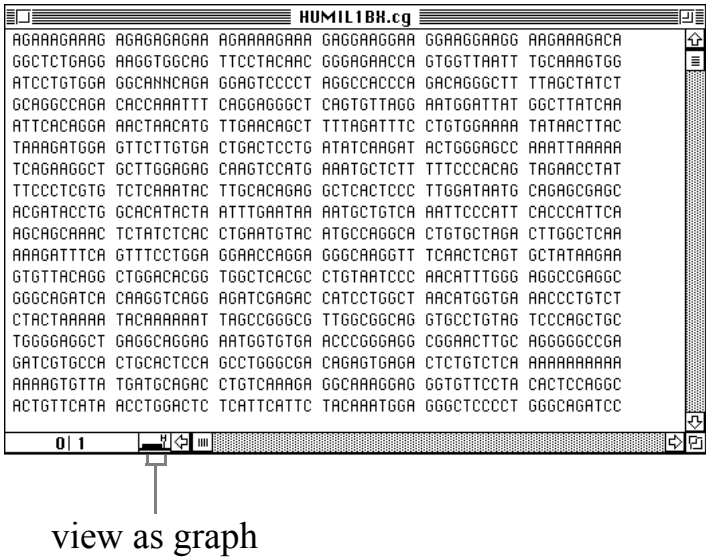


With CAD Gene™, you can view sequences in two ways -- graphically or as nucleotides. You can easily switch between the two viewing options, just press the "View as nucleotide" control at the bottom of the sequence window.



Click on the "view as nucleotides" control. The window will reset to display the nucleotide sequence of the file.

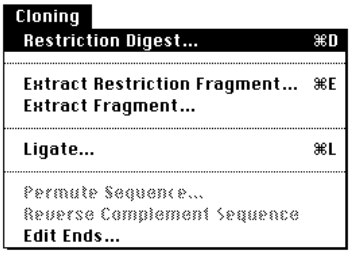
Tip: If a site on the graphic is selected, and the "view of nucleotide" control is selected, the sequence comprising the site will be selected in the nucleotide sequence. This is an easy way of verifying the sequence contents of sites.



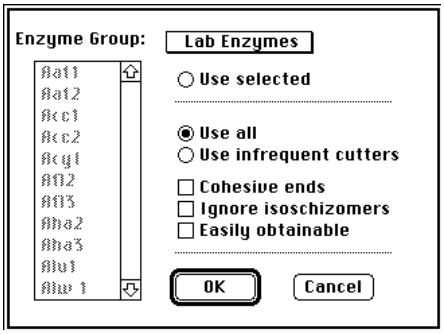
In the nucleotide view, the sequence can be edited as if it were a regular text file. CAD Gene™ will allow only the entry of valid sequence characters, and CAD Gene™ provides a number of

tools to aid in sequence entry and analysis. When the sequence is edited, the locations of all sites on the sequence are automatically updated.

Select the "view as graph" control to return the window to its initial appearance, then select " Restriction Digest" from the "Cloning" menu.

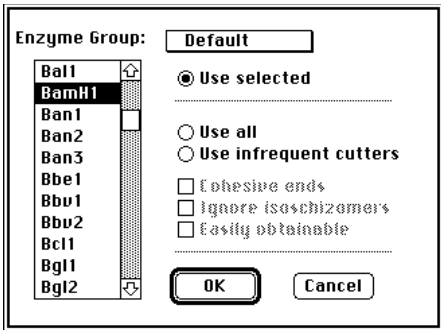


The "Select restriction enzymes" dialog box will appear.



CAD Gene™ provides for great flexibility in selecting which restriction enzymes to use in a restriction digest. CAD Gene™ allows for saving multiple restriction enzyme list, so, for example, you can create lists of enzymes that are frequently used by the lab and are available in lab stock, or create lists of enzymes available from a particular supplier.

Selecting "Use infrequent cutters" will use only enzymes that have a recognition sequence of 6 nucleotides or greater. "Use selected" allows for the selection of individual enzymes. For example, it is possible to check where "BamH1" and "Hind3" cut the sequence.

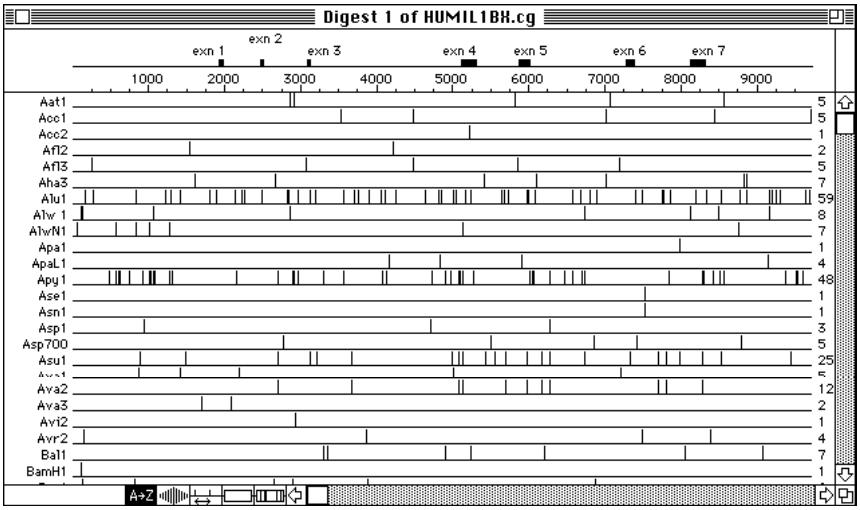


Note: To select multiple items in a list, hold the "shift key" down while making the selection and drag the cursor. To select discontinuous items, hold the "command" key down when selecting additional enzymes.

Tip: To scroll to a particular enzyme, you can just type its name.

In most cases, it is most useful to select "Use all." CAD Gene™ performs restriction digest extremely rapidly; to digest a 5000 nucleotide sequence with all enzymes takes less than 6 seconds an a Macintosh Classic.

Make sure "Use all" is selected, then select the "OK" button. The "Digest of…" window will appear.

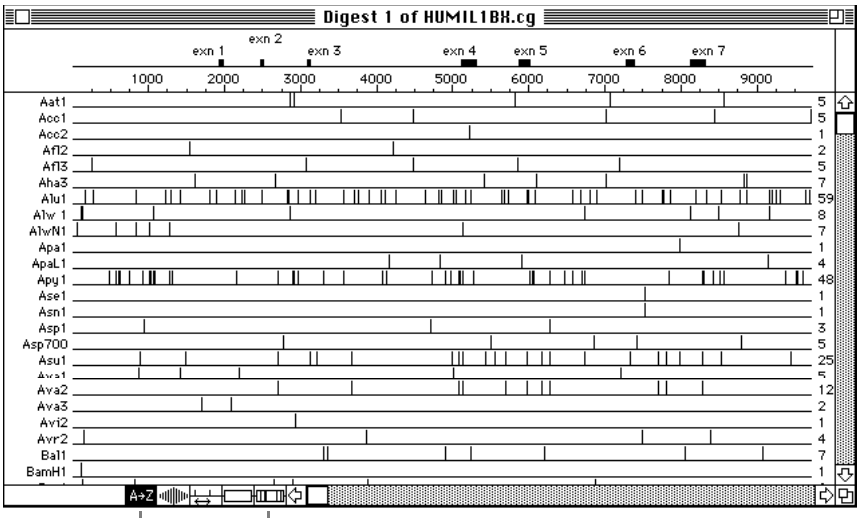


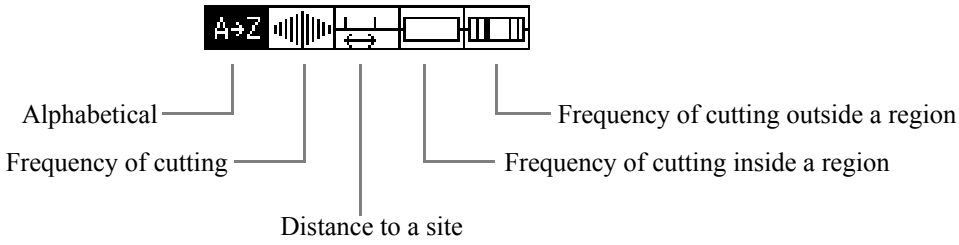
When the restriction site list appears, it is initially sorted alphabetically by name.

Tip: Typing a restriction enzyme's name while the restriction digest window is foremost will scroll that enzyme's site record into view.

A unique feature of CAD Gene™ is its ability to interactively sort the restriction enzyme list. By sorting the restriction enzyme list, it is possible to answer such questions as: "Which enzymes cut near location 500, but do not cut the sequence elsewhere?", or "Which enzymes do not cut within exon one, but cut near its borders?"

Sorting is accomplished by using the "Sort" pallet at the bottom of the digest window, or by selecting "Set Sort criteria…" from the "View" menu. In practice, it is usually more convenient to sort the enzyme list using the "Sort" pallet.



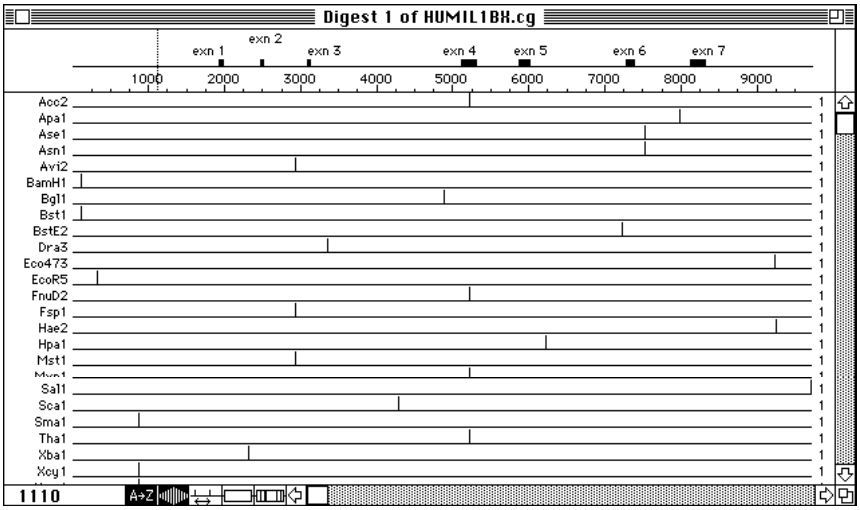


Sorting is accomplished by selecting one of the tools, and then optionally selecting the site(s).

Note: To select sequential sort tools, hold the shift key down when selecting tools.

Tip: Holding the option key down while selecting a sort tool reverses the order of the sort (increasing order of frequency becomes decreasing order of frequency).

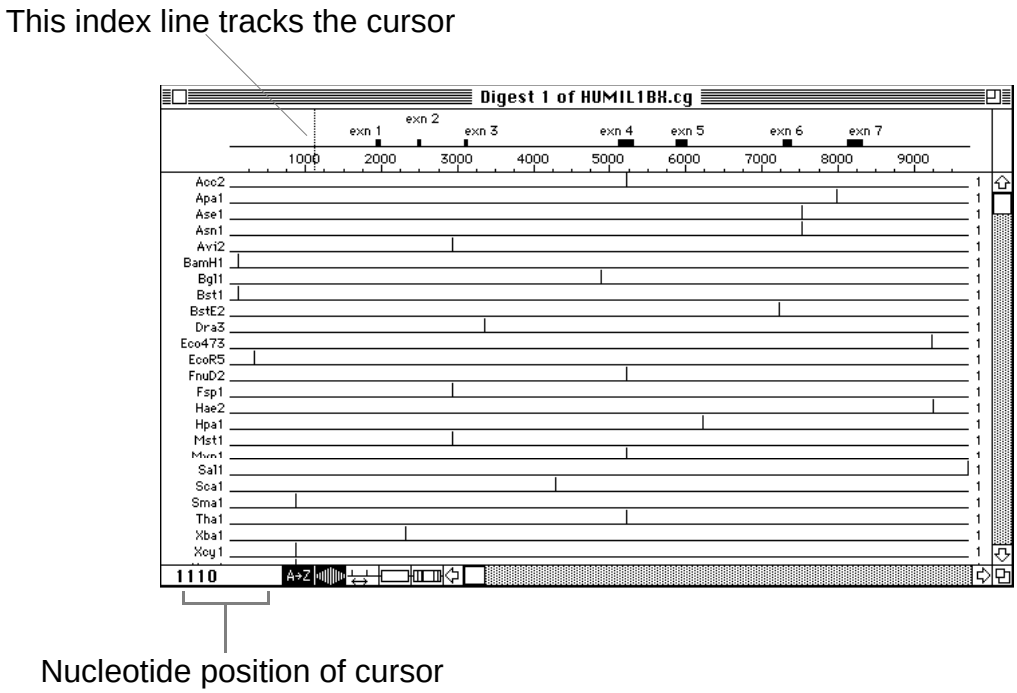
Select the "Frequency of cutting" tool at the bottom of the digest window.



The order that the restriction enzymes are now displayed in increasing order of frequency of cutting the sequence: those that cut the sequence once are displayed first. Enzymes that do not cut the sequence are not displayed. The list of enzymes that do not cut the sequence can be viewed by selecting "Display No Cutters" from the "Cloning" menu.

Next, while holding the shift key down, select the "Distance to a site" tool from the sort pallet. You *must* hold the shift key when selecting the "Distance to a site" tool from the sort pallet; if you do not, the "Sort by frequency of cutting" will be lost. To sort by "distance to a site", a starting location is selected on the sequence, and the distance from this site to the nearest cut site is computed for each restriction enzyme. These distances are then used to sort the restriction enzyme list in order of increasing distance.

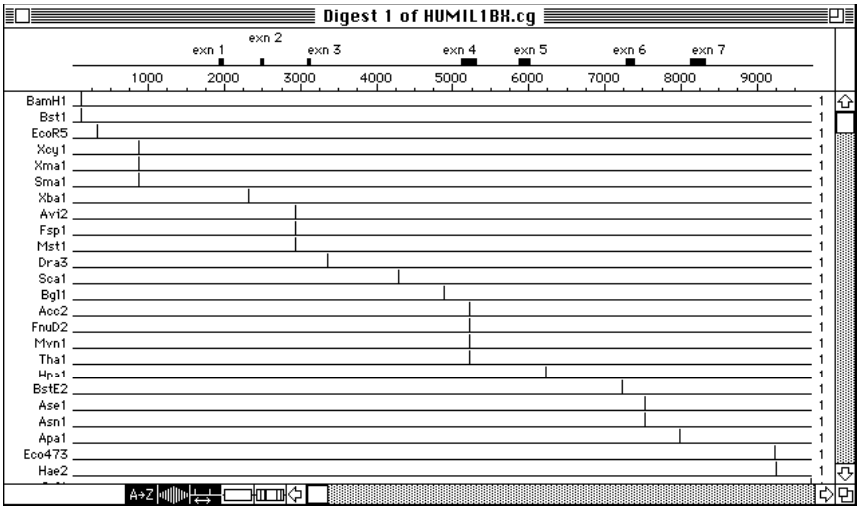
After you select the "Distance to a site" tool, notice that the "Distance to a site" tool on the pallet is highlighted, and that the cursor, when within the limits of the restriction digest window, is now a cross-hair (+). Now you must select the site that will be used for the distance calculation for sorting. This is done by moving the cursor to the location that is to be used to calculate the distances, and depressing the mouse button. When the cursor is over the restriction digest, the nucleotide position of the cursor is displayed in the lower left corner of the restriction digest window, and an index line tracks the cursor over the graphic representation of the sequence.



For sorting, sites can also be selected by clicking the picture of a region or site in the graphical representation of a sequence at the top of the digest window. When regions and sites are selected

on the graphic of the sequence, the locations of the end point of the region are displayed in the lower left corner of the window.

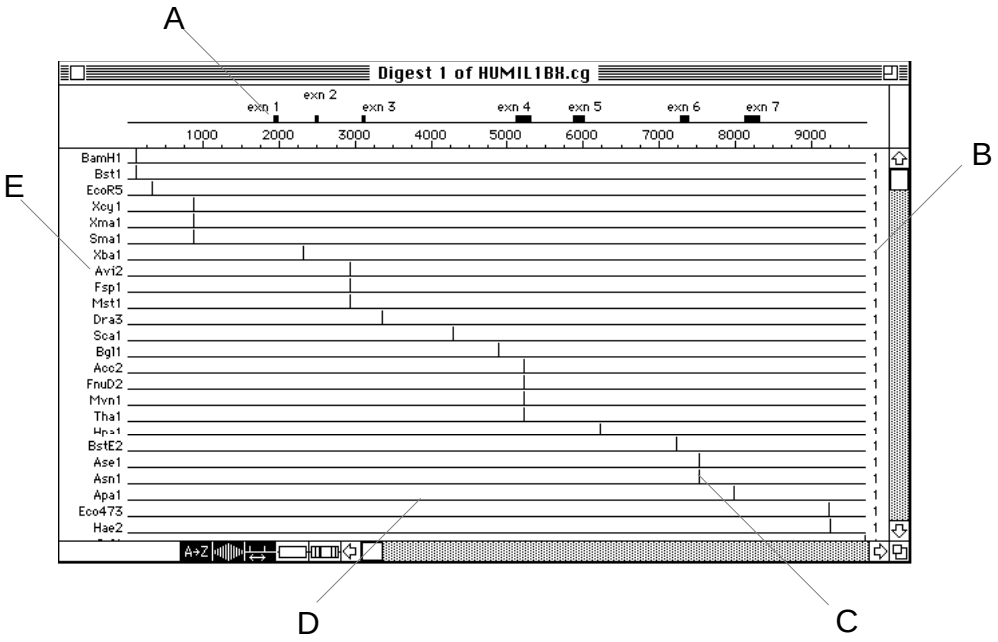
For this exercise, we just want to see where the unique (single site) restriction sites are located. To do this most easily, it is simplest to sort the enzyme list by distance of the site from the 5' end of the sequence. After the "Distance to a site" tool has been selected (with the shift key down), move the cursor to the left side of the digest window and press the mouse button. The number '1' should be in the lower left corner of the digest window. When the mouse button is released, the enzyme list will be sorted by frequency of cutting and distance to the 5' end of the sequence.



Notice that the restriction enzymes are now also sorted in order of increasing order of distance from the start of the sequence. Actually, the restriction enzyme list has been sorted by three criteria: It was first sorted by number of cut sites, then by distance to the start of the sequence and finally alphabetically by restriction enzyme name.

Tip: *In our experience, it is most useful to print out a hard copy of the restriction digest sorted by number of cut sites and distance to the start.*

The restriction digest window is designed to allow easy access to all the information necessary to plan a cloning project.



A. Clicking on the graphical representation of a region (or point site) will display the location (in nucleotides) of the regions endpoints, in the lower left corner of the window. Drop lines also appear over the restriction site list to help determine the relative location of the region's endpoints and the restriction enzyme cut locations.

B. Clicking on the number of cut sites will display the recognition site for the restriction enzyme (useful to determine if two enzymes leave compatible overhangs).

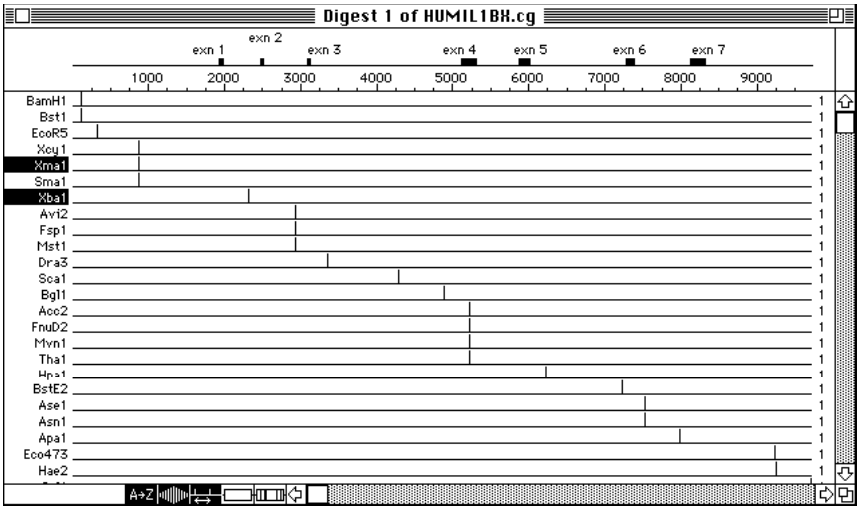
C. Clicking on a site marker will display the location of this restriction enzyme cut site.

D. Clicking on fragment will display the location (in nucleotides) of the fragment's endpoints, in the lower left corner of the window. Above the cursor will appear the length of the fragment, and the order of the fragment (A - largest to Z - smallest). Double clicking on a fragment will extract the fragment -- the fragment will appear in its own window as a new sequence. This is the quickest way of extracting a fragment produced by a single enzyme digest.

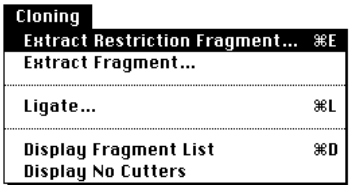
E. Clicking on the enzyme name selects the enzyme. Enzymes selected can be used in performing combined restriction digests (by selecting "Display Fragment List" from the "Cloning" menu) or for extracting restriction fragments from combined restriction digests (by selecting "Extract Restriction Fragment…" from the "Cloning" menu). To select multiple enzymes, hold the "command" key down when selecting enzymes.

Tip: To mark an enzyme for a partial restriction digest, hold the "option" key down when selecting an enzyme. Any combination of complete/partial digests with multiple enzymes is possible with the appropriate use of the "command" and "option" keys. Caution: The number of fragments produced by partial restriction digests increases exponentially with the number of cut sites.

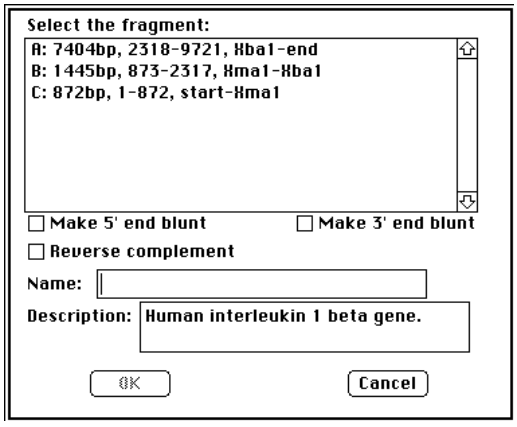
Select the enzymes “Xma1” and “Xba1” by clicking on the enzyme names with the “command” key depressed (to allow for multiple, discontinuous selections).



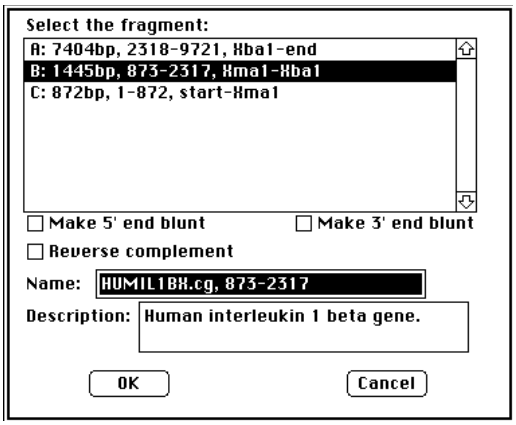
Next select “Extract Restriction Fragment…” from the “Cloning” menu.



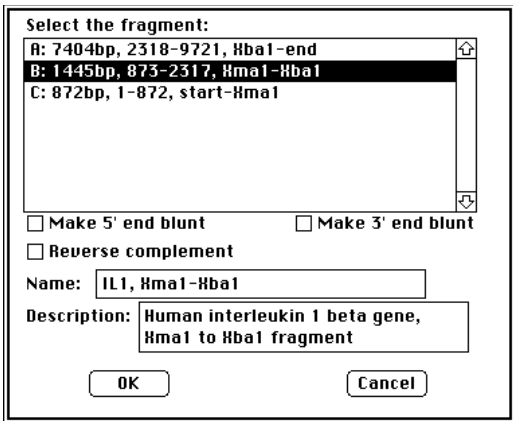
The “Extract restriction fragment” dialog will appear, presenting a list of restriction fragments produced by the selected enzymes.



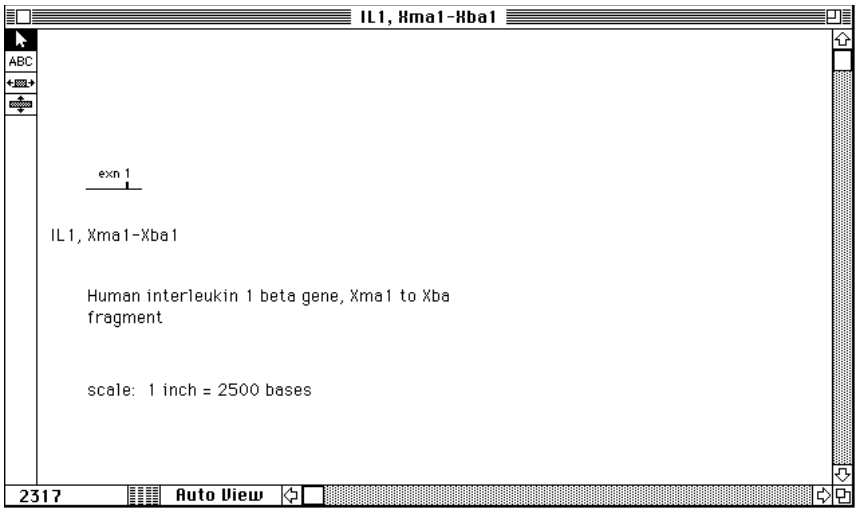
Select the fragment “B: 1445bp, 873-2317, Xma1-Xba1”.



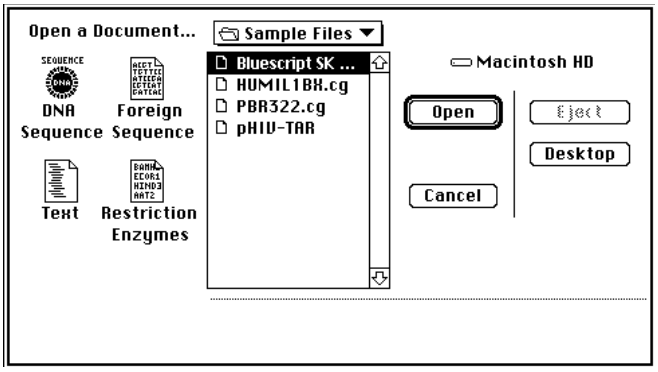
The "Name" and "Description" fields can be edited as desired; in this case they have been edited to reflect the restriction enzymes used to create the fragment.



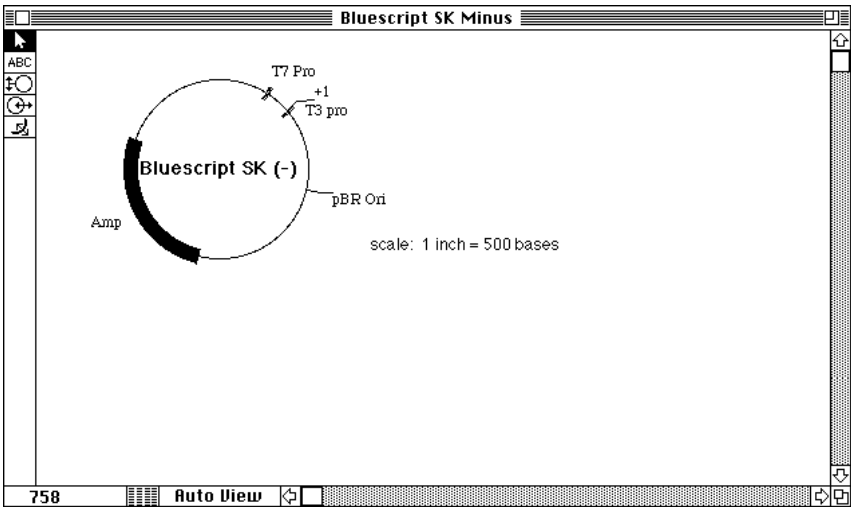
When done, select the "OK" button. The new sequence "IL1, Xma1-Xba1" will appear.



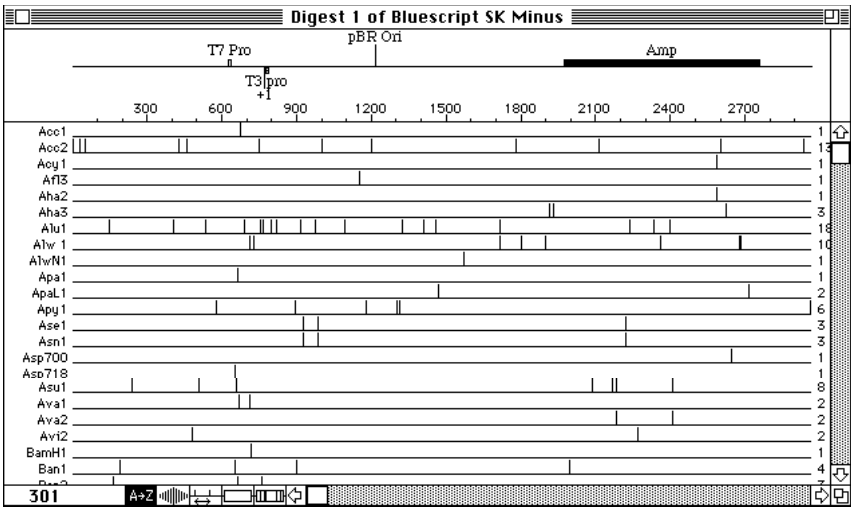
This completes generation of the first fragment to be used in this example. The next step is to open and extract the cloning vector fragment. Select “Open” from the “File” menu and select the sequence file “Bluescript SK Minus”.



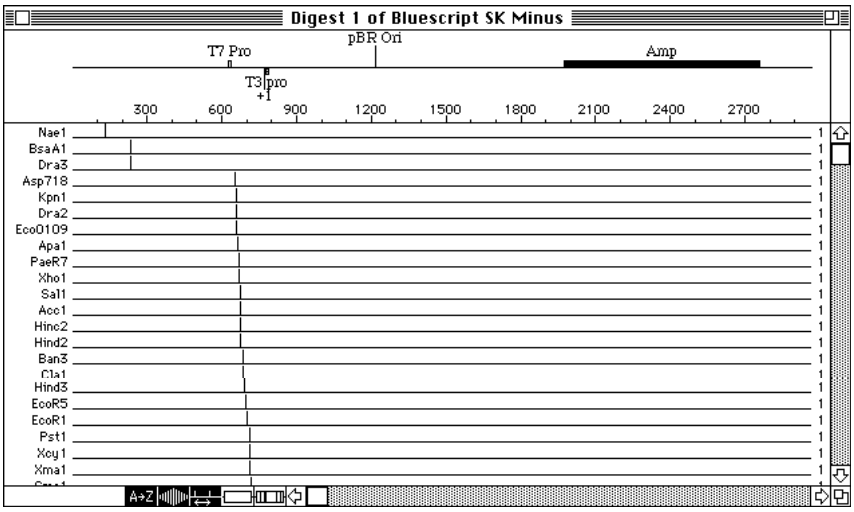
Select the “Open” button.



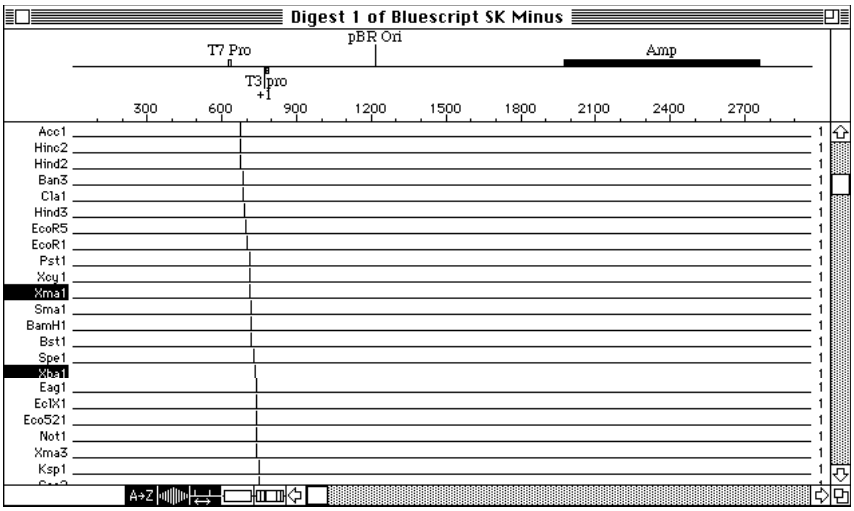
Select “Restriction Digest” from the “Cloning” menu, select “Use all” in the “Restriction digest” dialog and select the “OK” button.



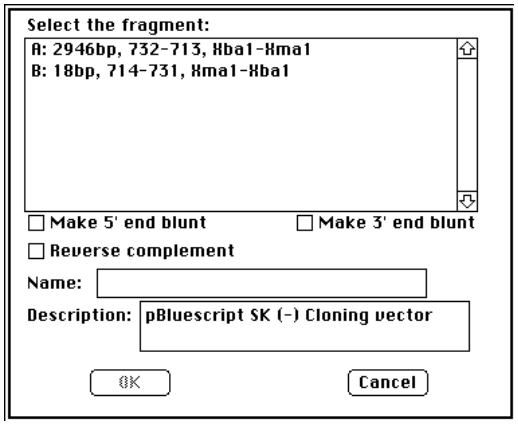
Again, select the "Frequency" tool from the "Sort" pallet, then select the "Distance to a site" tool (with the "shift" key down), and select the 5' end of the sequence. This is done in a manner identical to as was done to with the IL1.



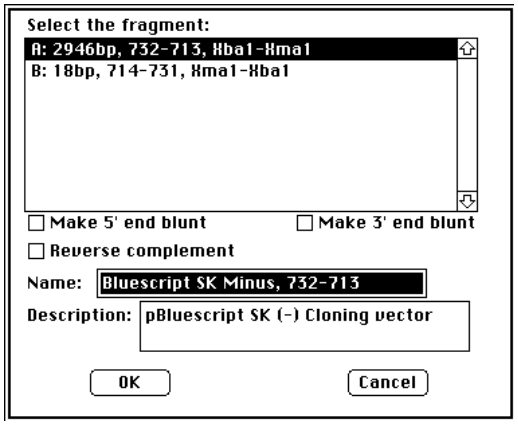
Scroll the restriction enzyme list (using the vertical scroll bar) until the enzymes "Xba1" and "Xma1" are visible in the restriction digest window. Select "Xba1" and "Xma1" by clicking on their names (again, holding the "command" key down).



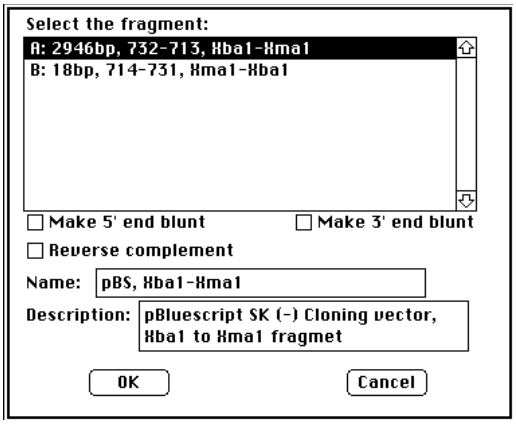
Select “Extract Restriction Fragment…” from the “Cloning” menu.



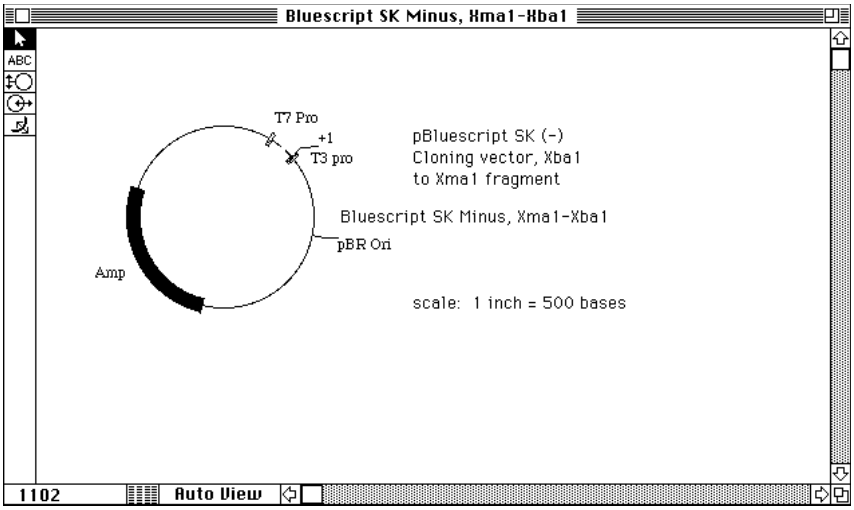
Select the fragment “A: 2946bp, 732-713, Xba1-Xma1“.



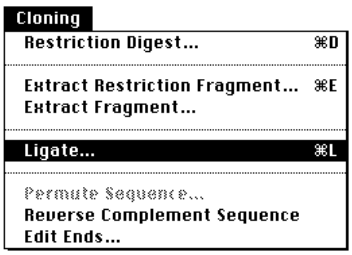
Edit the “Name” and “Description” fields.



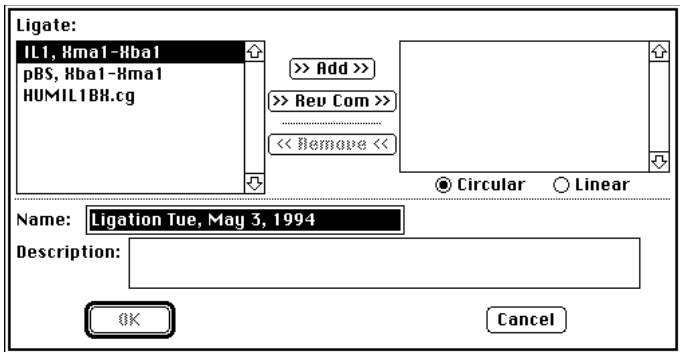
Select the "OK" button.



Now the IL1 exon 1 fragment and the Bluescript fragments have been generated, we can produce the final product. Select "Ligate…" form the "Cloning" menu.

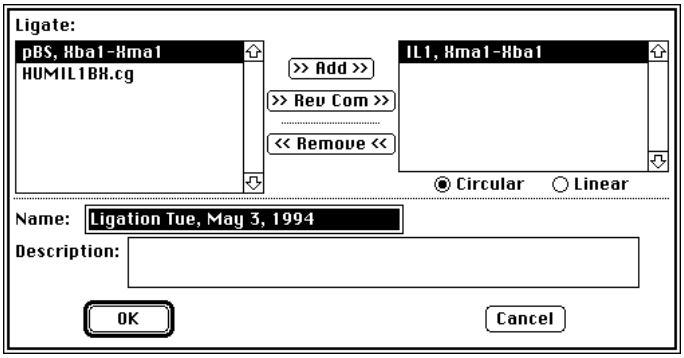


The "Ligate" dialog will appear.

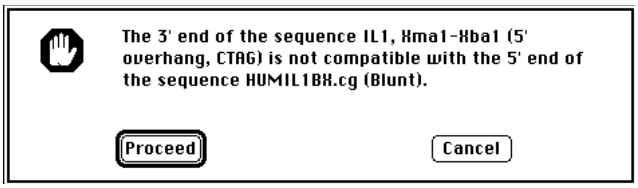


The "Ligate" dialog function similar to FontDA mover from Apple® Computer. The left hand list contains all currently open, linear sequences. Sequences are selected from the left hand list, and moved to the right hand list (the ligation product) using the "Add" and "Rev Com" buttons. The "Rev Com" button adds the reverse complement of the sequence selected in the left hand list to the ligation product being constructed in the right hand list. The selected items can be removed from the right hand list using the "Remove" button.

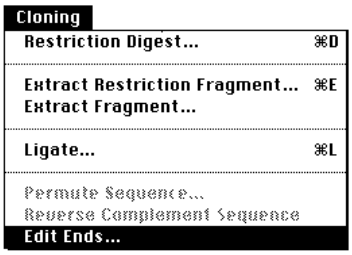
Add the sequence "IL1, Xma1-Xba1" to the ligation list by selecting the sequence name in the left hand list and selecting the "Add" button.



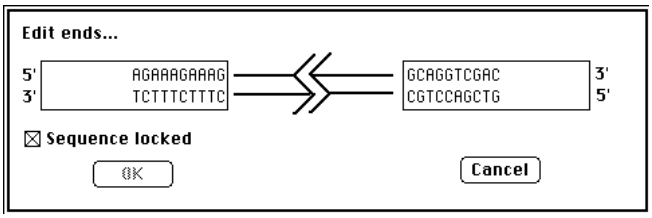
CAD Gene™ intelligently checks the ligation product as it is built. If you add a sequence that contains incompatible overhangs, a warning alert box will appear. The warning alert details the nature of the incompatibility.



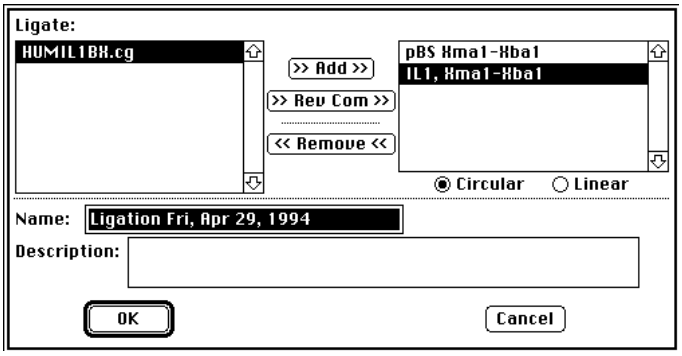
To easily double check the nature of the ends of a sequence, when a linear sequence window is front most, there is an "Edit Ends…" utility in the "Cloning" menu.



The "Edit ends" dialog allows you an easy method of viewing (as double stranded DNA) and editing the ends of a linear DNA fragment.

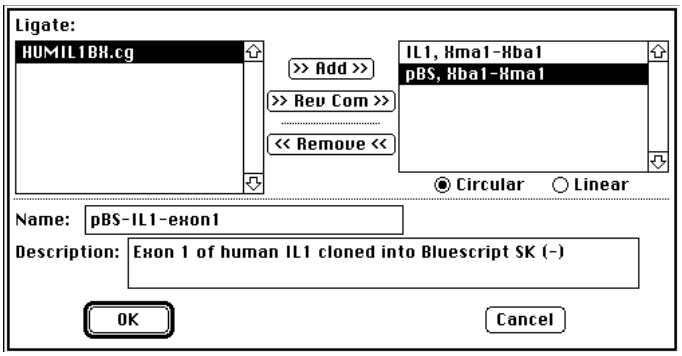


Add the sequence "pBS, Xba1-Xma1" to the ligation list by selecting the sequence name in the left hand list and selecting the "Add" button.

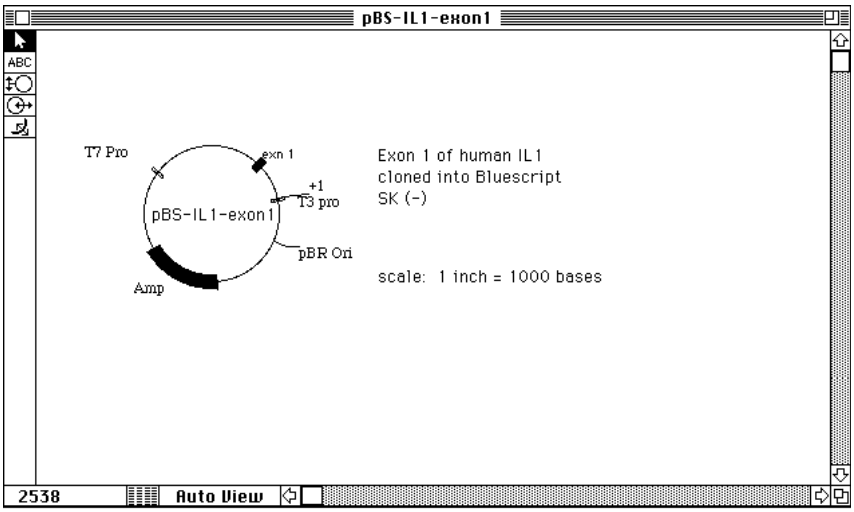


A default name for the new sequence produced by the ligation is provided; it is usually better to provide a more informative name, here we've renamed the sequence "pBS-IL1-exon1". Just type this string into the "Name" item of the dialog. This name will both appear as the title of the new window, and on the graphic of the sequence.

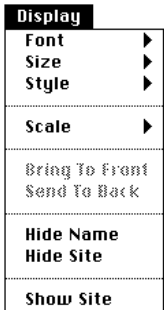
Also associated with each sequence is a "Description". This is a slightly longer text string, associated with the sequence, that allows for a slightly more detailed description of what the sequence contains. Enter the text "Exon 1 of human IL1 cloned into Bluescript SK (-)" into the "Description" box of the "Ligation" dialog. When you are done, select the "OK" button (or hit the "Return" key.



The “pBS-IL1-exon1” window will appear.



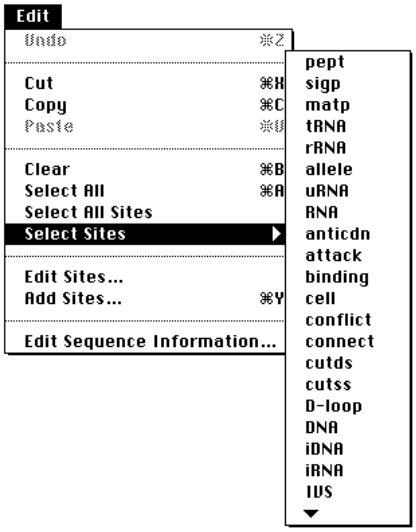
Congratulations! You have just successfully completed a simple, sample cloning project, using the CAD Gene™ software. Hopefully, this exercise has helped you to become familiar with some of the features of the CAD Gene™ software. To fully appreciate the advantages of the CAD Gene™ program design, we would encourage you to begin using CAD Gene™ for your own cloning projects.



CAD Gene™ faithfully calculates the location of all the sites on the resultant sequence. In addition, CAD Gene™ also remembers the source of all sequences that are produced by ligation by automatically adding “Genealogy” sites to sequences that are produced by extracting fragments.

CAD Gene™ provides numerous utilities for viewing and manipulating sites on sequences. The name of sites and regions can be edited on the screen with the “Text” tool on the tool pallet. Sites can be selected, singly or in combination, with

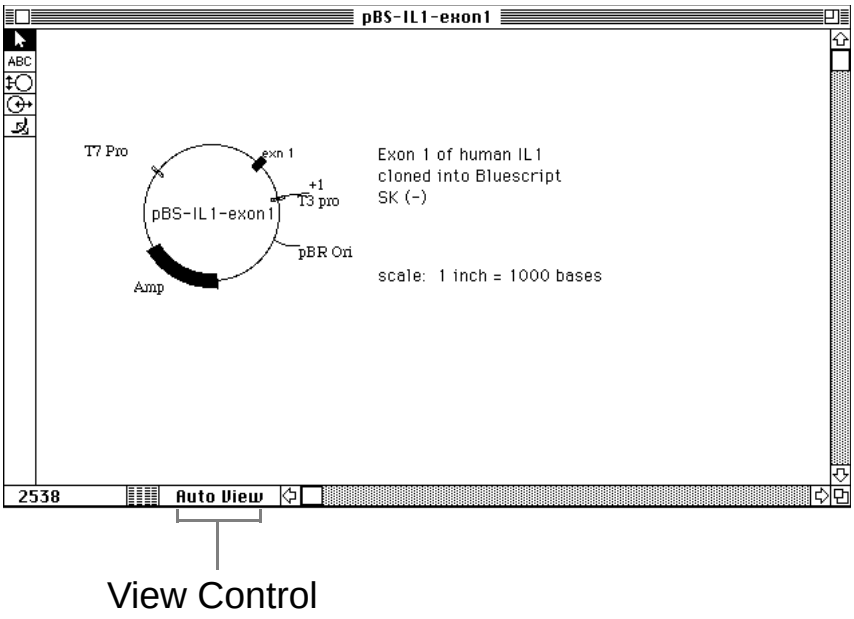
the “Pointer” tool from the tool pallet. Selected sites can be manipulated -- the name made invisible, the site made invisible, the font size/style/face changed, etc. These commands are located on the “Display” menu.



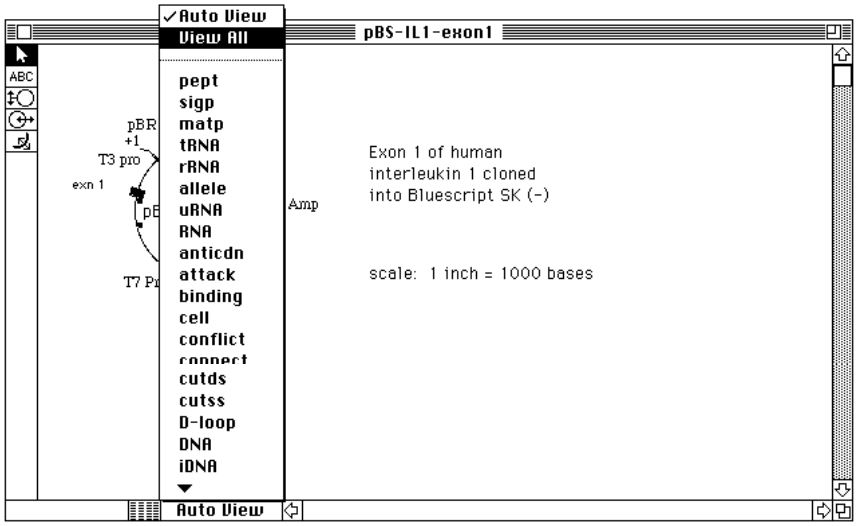
CAD Gene™ also provides a number of tools for manipulating groups of sites. Sites can be selected singly, using the “Pointer” tool from the tool pallet, or all the sites can be selected, by selecting “Select All Sites” from the “Edit” menu. Sites can also be selected by group by selecting “Select Sites” from the “Edit” menu.

Tip: *One of the most powerful features of CAD Gene™ is its ability to copy sites between sequences. A common problem that arises is that sites will be annotated on some sequences, but not be annotated on others. When a site is copied from the graphic display of a sequence, it can be pasted into the graphic display of another sequence. For example, to transfer all the annotated sites from pBR322 to Bluescript (which contains a partial pBR backbone), just copy the sites from pBR322 and paste them into Bluescript. For those sites in pBR322 that are not present in Bluescript, a warning alert will be shown.*

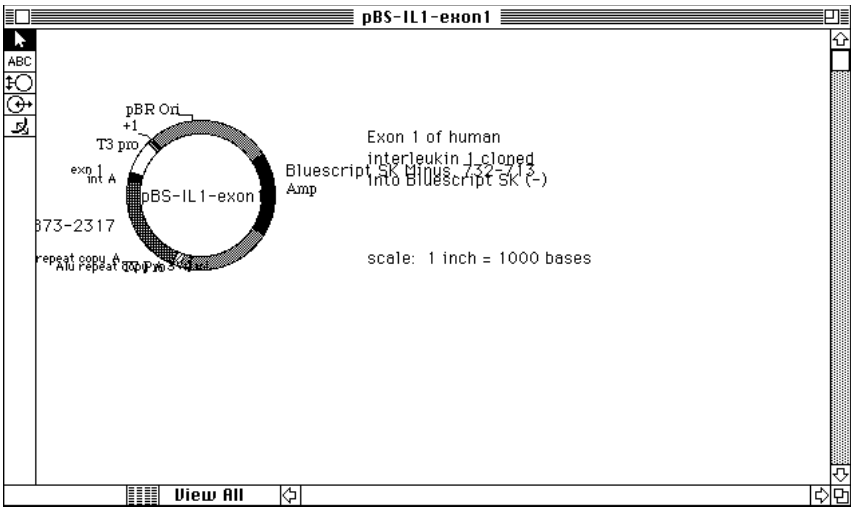
Most sequences contain too many sites to be easily displayed all at one time. To get around this problem, CAD Gene™ allows sites to be classified as to “type”. “Types” are defined by function of the region -- such as “pept”, “tRNA”, or “orgrpl”. Sites can be optionally flagged not to be displayed. If all the sites of interest were displayed at one time, the graphic of the sequence rapidly becomes so crowded that individual sites become difficult to discern. To demonstrate this problem, just select “View All” from the “View” control at the bottom of the “pBS-IL1-exon 1” sequence window.



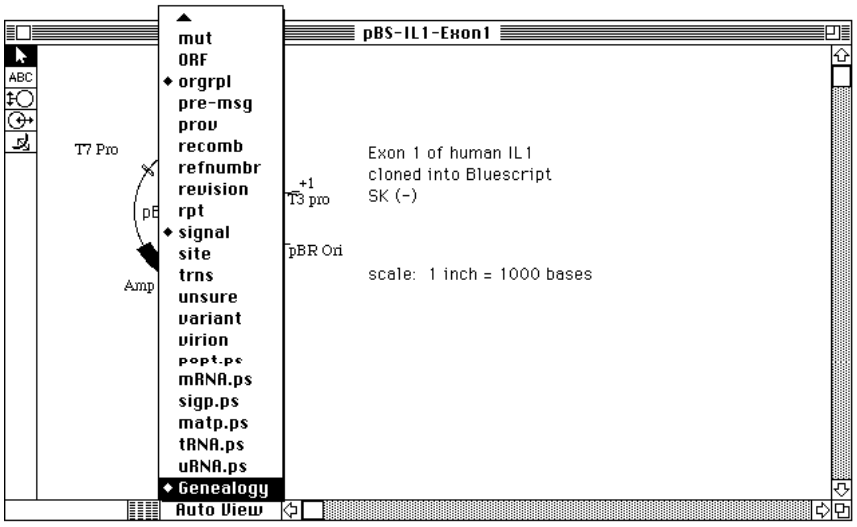
After clicking on the "View" control, a pop-up menu of all the site types will appear. To view all the sites on the sequence, select "View All" from the pop-up menu.



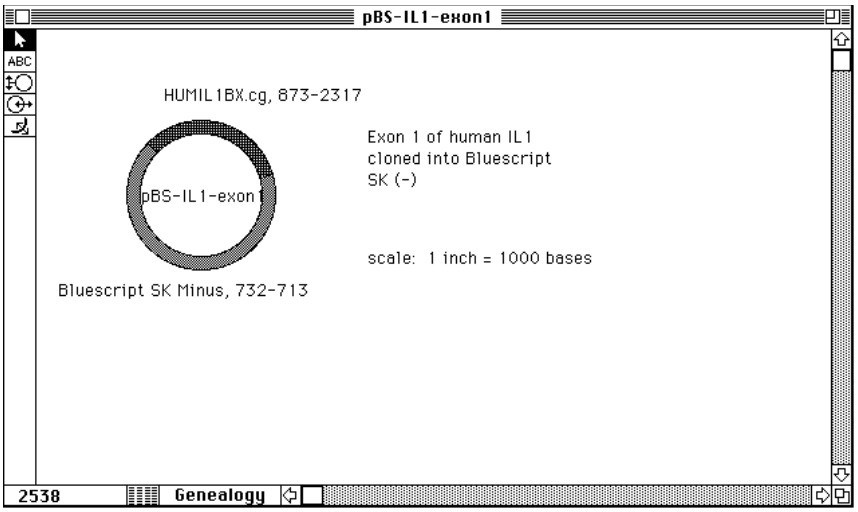
When "View All" is selected, the "Hide Site" option is over-ridden.



One can also view only selected groups of sites by selecting the type name in the “View” control menu. Select “Genealogy” from the “View” control menu.



The “genealogy” sites will be displayed.



The “genealogy” sites displayed here are automatically generated whenever a fragment is “extracted” from a restriction digest, and are a simple, direct mechanism of unambiguously tracing the origin of the nucleotides in the sequence.