# Intersect

Version 1.0

## User's Manual

(last revision: 12-18-02)

© 2001 University of California

# Table of Contents

1. Introduction 3

2. Installation 4

3. Using Intersect 5

- Adding sets 5
- Adding files to a set 6
- Removing files from a set 7
- Adding names to a set 7
- Changing set colors 7
- Analyzing the data 8
- Displaying Venn diagrams 9
- Displaying outlines 9
- Displaying set labels 9
- Displaying intersection information 10
- Filtering intersection information 11
- Saving set/intersection information 11
- Printing the Venn diagram 11
- Saving the session 12
- Opening a saved session 12

Appendix A: Session file format 13

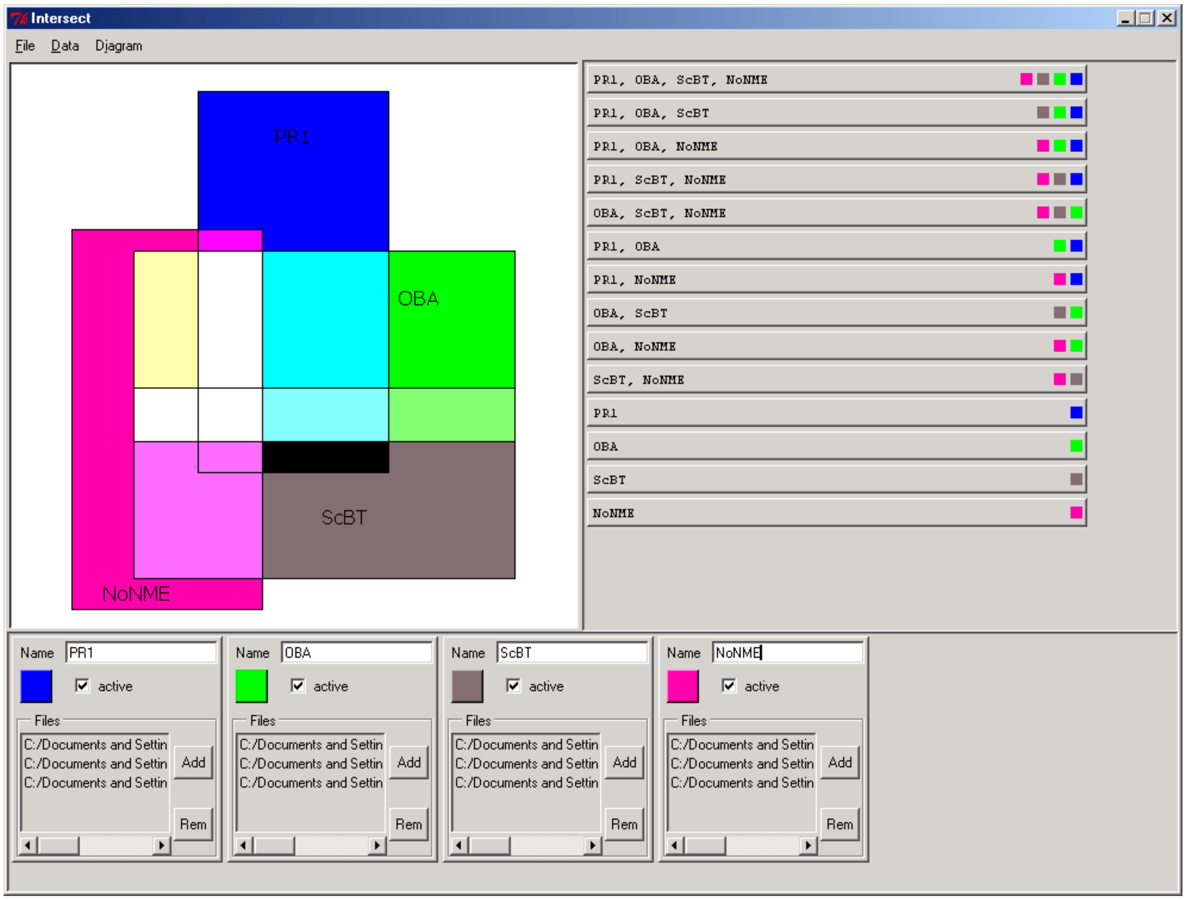
Appendix B: Known bugs 14

Appendix C: Frequently asked questions 15

# 1. Introduction

Welcome to Intersect, version 1.0.  If you're reading this section, you're probably wondering what this program is all about. In the broadest sense, Intersect is a program for biological sequence analysis. More precisely, it is a post-processing program for sets of output files generated by sequence database searches. Intersect identifies sequences reported across sets of output files and displays this information visually.

Here's an example of what the Intersect program looks like:



```
PR1, OBA, ScBT, NoNME   🟪🟫🟩🟦
PR1, OBA, ScBT          🟫🟩🟦
PR1, OBA, NoNME         🟪🟩🟦
PR1, ScBT, NoNME        🟪🟫🟦
OBA, ScBT, NoNME        🟪🟫🟩
PR1, OBA                🟩🟦
PR1, NoNME              🟪🟦
OBA, ScBT               🟫🟩
OBA, NoNME              🟪🟩
ScBT, NoNME             🟪🟫
PR1                     🟦
OBA                     🟩
ScBT                    🟫
NoNME                   🟪
```

Name PR1 🟦 ☑ active
C:/Documents and Settin
C:/Documents and Settin
C:/Documents and Settin

Name OBA 🟩 ☑ active
C:/Documents and Settin
C:/Documents and Settin
C:/Documents and Settin

Name ScBT 🟫 ☑ active
C:/Documents and Settin
C:/Documents and Settin
C:/Documents and Settin

Name NoNME 🟪 ☑ active
C:/Documents and Settin
C:/Documents and Settin
C:/Documents and Settin

The Intersect program allows users to explore the interconnections between sets of database search results. When each set of search result files is created by searches using query sequences that are related, Intersect can show the relationships between the sets of query sequences. The Intersect display gives an intuitive view of where the database search results have reported the same sequence for searches involving different sets of queries. These sequences are often of biological interest, as they may represent evolutionary linkages between sequence families and superfamilies.

© 2001 University of California

# 2. Installation

Intersect is written in Python (version 2.1), making it platform independent as long as Python is installed. The first step in installing Intersect is thus the installation of Python, the most recent versions of which are available at www.python.org. The Python installation process is quick and painless.

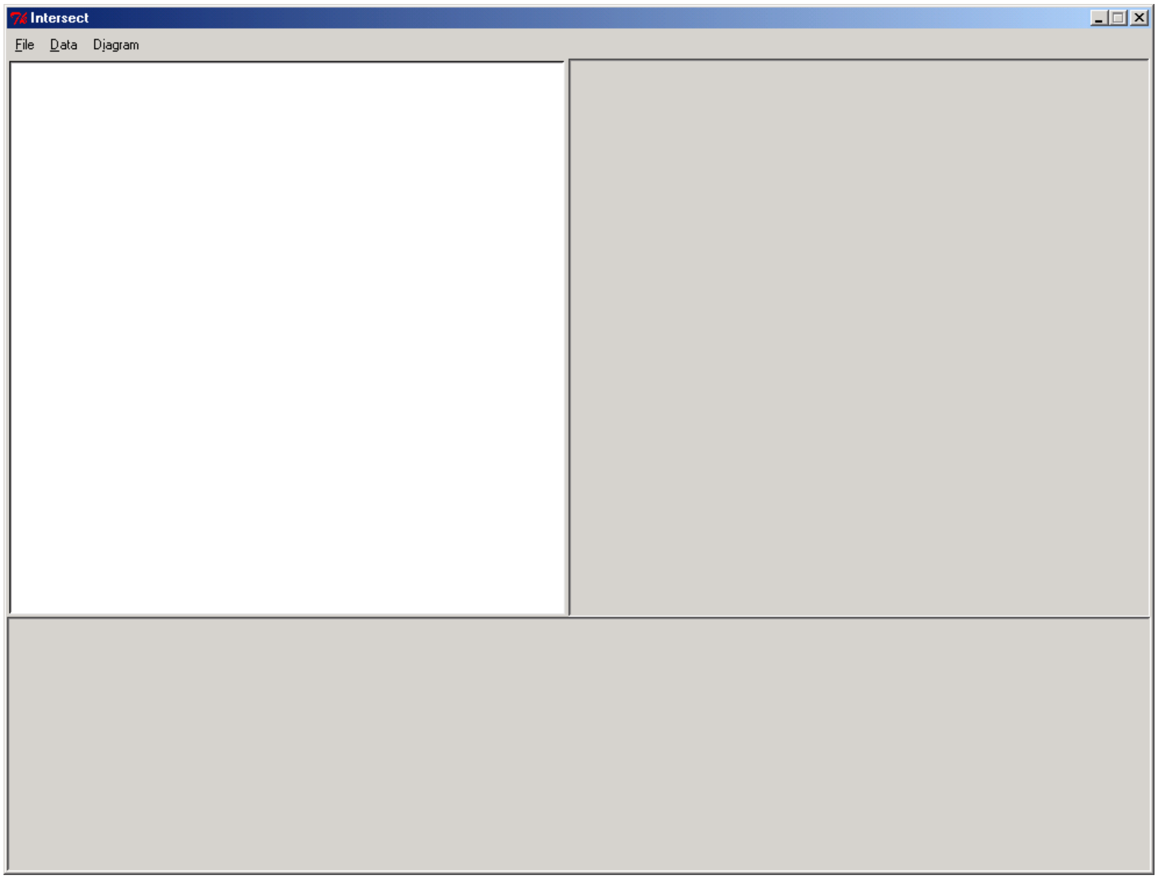
Intersect uses the Python Mega Widget objects library (Pmw), which is not currently distributed with Python. This library is available via SourceForge at http://sourceforge.net/projects/pmw, and its installation is also quick and painless. Be sure to install the Pmw library within the top level of the Python distribution directory so that Python can find it at run time.

*Note: If you have already installed the Chimera program (from www.cgl.ucsf.edu) then both Python and the Pmw library are installed, but you'll need to copy the Pmw directory from the Chimera directory to the Python directory in order for Python to have access to it.*

Installing Intersect itself is very easy—just copy the file Intersect.py to your computer. You can run the program from a command prompt by typing its name, or via a window manager by double clicking on it.

# 3. Using Intersect

The Intersect program can be started by either typing “Intersect.py” on a command line, or by double-clicking on the file *Intersect.py*. This will bring up the graphical interface:



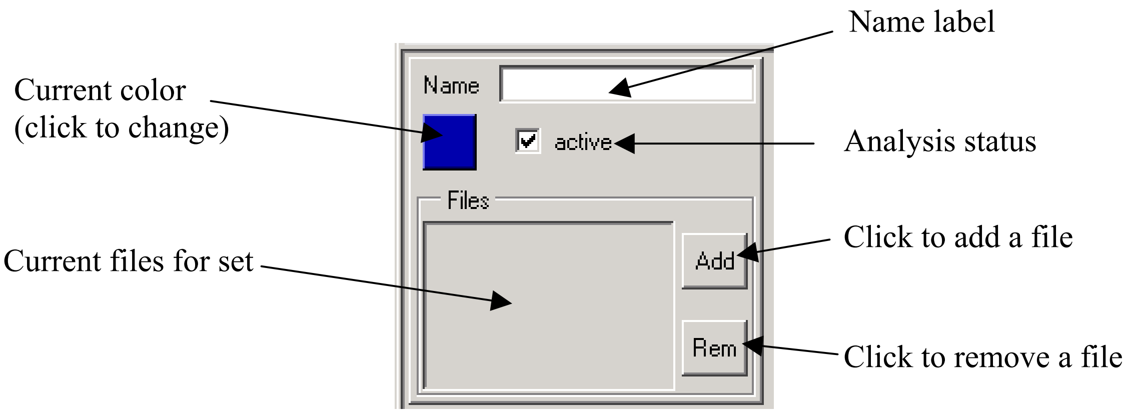
It looks pretty boring at first, but don’t worry, you’ll be adding lots of great data and color. First, however, note that there are three major panels. The upper left panel is a display which is used to draw Venn diagrams. The upper right panel displays basic set overlap information. The lower panel houses the controls for each set of database search files.

## Adding sets

The first step in adding a set of database search files is to click on “Add Set” from the “Data” menu at the top of the interface (alternatively, you can type Ctrl-N). This will add a new set of controls to the lower panel of the interface. This panel allows you to name the set, add or remove files from the set, view and change it’s representative color, and choose whether it’s active in the current analysis.

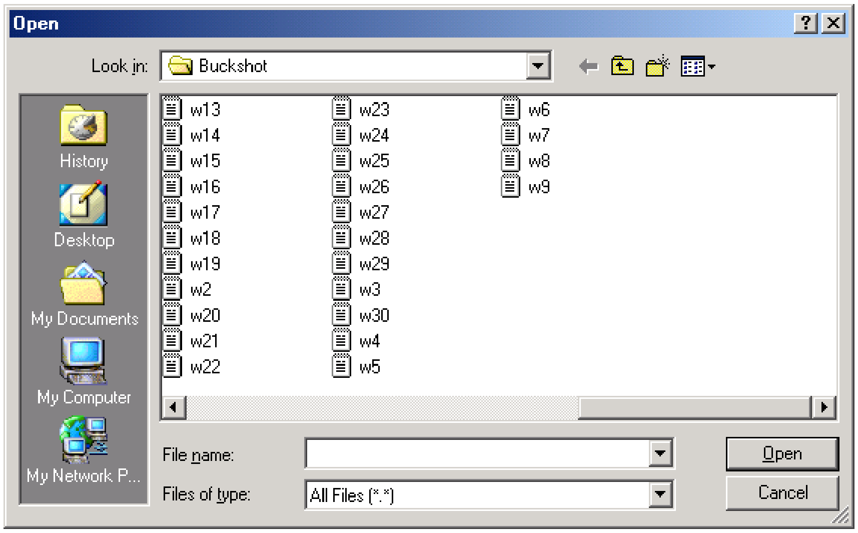
© 2001 University of California

Here’s what the controls look like:



## Adding files to a set

To add a file to a set, click on the set’s “Add” button. This will bring up a standard file browsing window from which to choose a file.



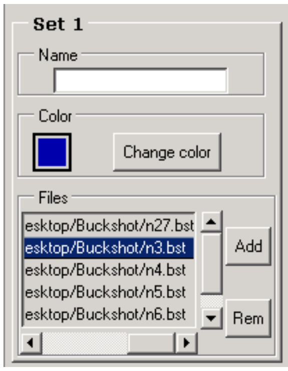
Just choose a file and click “Open”. The file will then show up in the set’s list of current files.

Intersect currently handles files from the popular database search programs, BLAST (both the NCBI and Washington University in St. Louis versions), FASTA, and PSI-BLAST.

© 2001 University of California

## Removing files from a set

To remove files from a set, select the file to remove by clicking on it in the set's list of current files.



esktop/Buckshot/n27.bst
esktop/Buckshot/n3.bst
esktop/Buckshot/n4.bst
esktop/Buckshot/n5.bst
esktop/Buckshot/n6.bst

Simply click the "Rem" button from the set's control panel and the file will be removed.

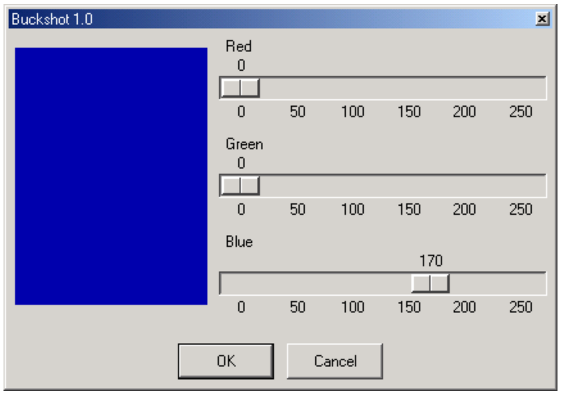
## Adding names to sets

Each set can be given a name. This name is displayed as a text label on the Venn diagram, and is used in the naming of sets in the text output (see Displaying set/intersection information) and in the set results panel. To name a set, simply type a name into the "name" entry box in the set's control panel.

Names can contain spaces and have no length restrictions, although exceeding the length displayed in the entry box of the control panel may cause the label to exceed the width of the display. In this case the display of the label will be truncated. If many sets are active, shorter names are more useful, since the results panel will only display 50 characters per set intersection (see "Analyzing the data" below).

## Changing set colors

To change the color used to represent the set in the display, click on the colored button in the set's control panel. This brings up a window with which to choose a new color.



Buckshot 1.0

Red 0

Green 0

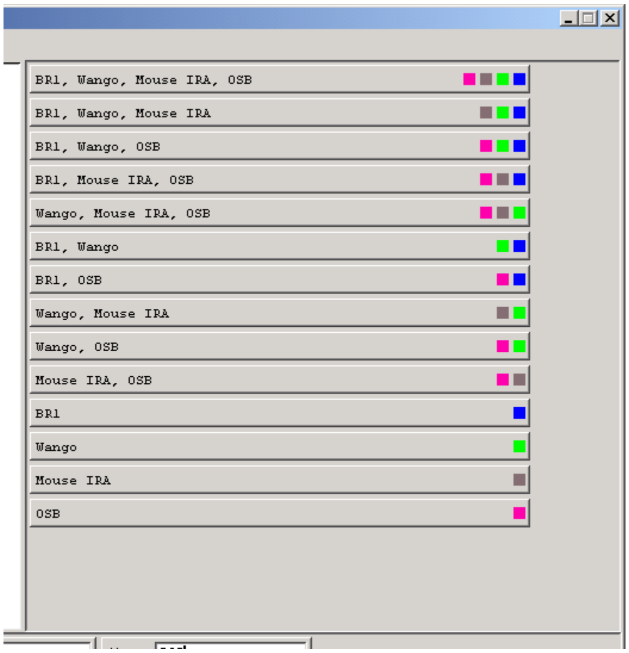
Blue 170

The current color is shown to the left, and will change as the sliders for red, green, and blue values are moved. Clicking the "OK" button will set the chosen color as the new set color.

*Note: In the Venn diagram display, set intersections for which there are no sequences are colored black. Choosing black as a set color is thus not recommended.*

## Analyzing the data

Once you have sets of database search files entered into their respective control panels, you're ready to analyze them for intersections. This is done by choosing "Update" from the "Data" menu (or alternatively by typing Ctrl-U). Intersect then parses each of the database search files and determines each non-empty overlap between sets of search files. These sets are listed in the results panel to the upper right:



```
BR1, Wango, Mouse IRA, OSB
BR1, Wango, Mouse IRA
BR1, Wango, OSB
BR1, Mouse IRA, OSB
Wango, Mouse IRA, OSB
BR1, Wango
BR1, OSB
Wango, Mouse IRA
Wango, OSB
Mouse IRA, OSB
BR1
Wango
Mouse IRA
OSB
```

Each horizontal bar represents a non-empty intersection between sets of database search results. The name of each set in the intersection is listed (up to 50 total characters) and the representative color of each set is shown at the right of the bar. Clicking on a horizontal bar brings up a window displaying the contents of the intersection (see “Displaying intersection information” below).

## Displaying Venn diagrams

To display a Venn diagram for the currently active sets, choose “Draw Venn” from the “Display” menu (or alternatively type Ctrl-V). A Venn diagram is only drawn if four or fewer sets are active. (Complete Venn diagrams for five or more sets contain too many regions for effective visualization, and often require irregularly shaped regions.) After making any changes to a set’s name or color, choose “Draw Venn” again to render those changes in the display.

The sets in the resulting diagram are colored according to the colors chosen in the set control panels. The intersections of the sets are colored by adding the colors of the sets. Intersections which do not contain any sequences are colored black.

*Note: The areas represented by the sets and their intersections do not represent their sizes (ie. number of sequences contained within). When looking at the display, keep in mind the phrase “Figure not drawn to scale”.*

## Displaying outlines

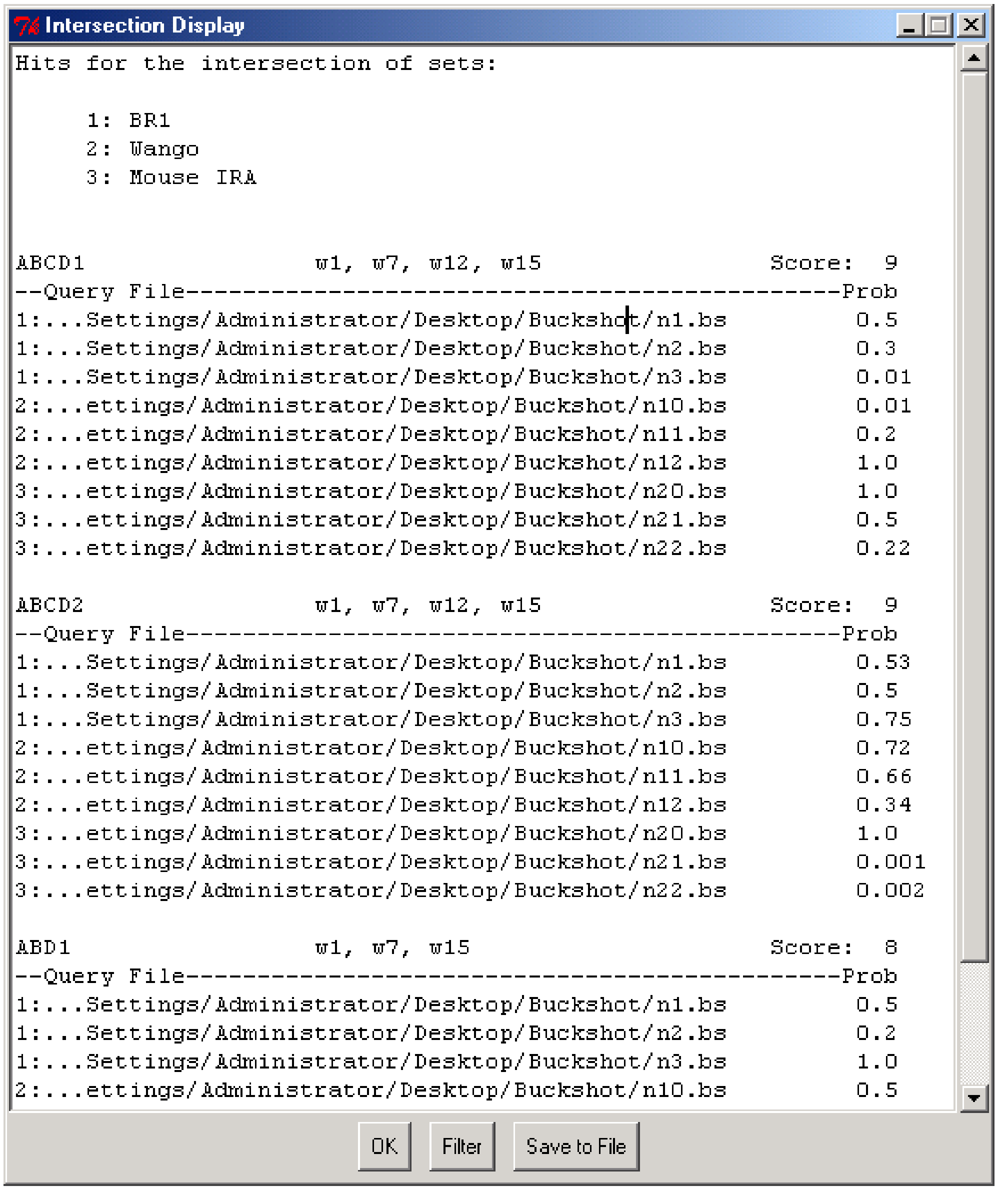
Black lines outlining the set boundaries in the display can be drawn by choosing “Show Outlines” from the “Display” menu. By default, outlines are not drawn.

## Displaying set labels

The name of each set can be included in the Venn diagram by choosing “Show Set Labels” from the “Display” menu. By default, the name labels are not drawn.

# Displaying intersection information

Clicking on a horizontal bar from the results panel or on a set intersection from the Venn diagram will bring up a window with information about the sequences in that particular intersection, with the exception of the empty (colored black).



```
Hits for the intersection of sets:

     1: BR1
     2: Wango
     3: Mouse IRA


ABCD1                w1, w7, w12, w15                Score:  9
--Query File----------------------------------------------Prob
1:...Settings/Administrator/Desktop/Buckshot/n1.bs         0.5
1:...Settings/Administrator/Desktop/Buckshot/n2.bs         0.3
1:...Settings/Administrator/Desktop/Buckshot/n3.bs         0.01
2:...ettings/Administrator/Desktop/Buckshot/n10.bs         0.01
2:...ettings/Administrator/Desktop/Buckshot/n11.bs         0.2
2:...ettings/Administrator/Desktop/Buckshot/n12.bs         1.0
3:...ettings/Administrator/Desktop/Buckshot/n20.bs         1.0
3:...ettings/Administrator/Desktop/Buckshot/n21.bs         0.5
3:...ettings/Administrator/Desktop/Buckshot/n22.bs         0.22

ABCD2                w1, w7, w12, w15                Score:  9
--Query File----------------------------------------------Prob
1:...Settings/Administrator/Desktop/Buckshot/n1.bs         0.53
1:...Settings/Administrator/Desktop/Buckshot/n2.bs         0.5
1:...Settings/Administrator/Desktop/Buckshot/n3.bs         0.75
2:...ettings/Administrator/Desktop/Buckshot/n10.bs         0.72
2:...ettings/Administrator/Desktop/Buckshot/n11.bs         0.66
2:...ettings/Administrator/Desktop/Buckshot/n12.bs         0.34
3:...ettings/Administrator/Desktop/Buckshot/n20.bs         1.0
3:...ettings/Administrator/Desktop/Buckshot/n21.bs         0.001
3:...ettings/Administrator/Desktop/Buckshot/n22.bs         0.002

ABD1                 w1, w7, w15                     Score:  8
--Query File----------------------------------------------Prob
1:...Settings/Administrator/Desktop/Buckshot/n1.bs         0.5
1:...Settings/Administrator/Desktop/Buckshot/n2.bs         0.2
1:...Settings/Administrator/Desktop/Buckshot/n3.bs         1.0
2:...ettings/Administrator/Desktop/Buckshot/n10.bs         0.5
```

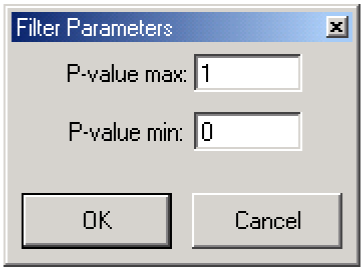
This display contains information concerning which files of each set reported each sequence. This information can be saved to a file by clicking the “Save to File” button.

*Note: The user can edit the text in this window, and the changes will be saved when the “Save to File” button is clicked on. This allows users to make notes directly in the data and save them for future reference.*

© 2001 University of California

## Filtering intersection information

At the bottom of the intersection display window (see image above), there's a "Filter" button. Clicking on this button gives you an opportunity to remove hits according to their score (P-value, or "Prob" as listed in the display). The "Filter" button brings up a small window in which you can input minimum and maximum values:



Filter Parameters

P-value max: 1

P-value min: 0

*Note: Filtering is not stochastic. In other words, if you filter once on a given set of minimum and maximum values, and then filter again with a different set of values, the results will be the same as if you filtered with the later parameters on the original intersection information. In other words, the filtering is not cumulative.*

## Saving set/intersection information

All of the information concerning the sets and their intersections (such as that displayed by clicking on the Venn diagram) can be saved at once by choosing "Save Clusterings" from the "File" menu. The user is prompted to provide a base filename, to which the set and intersection numbers are appended. (For example, if the user provides the base filename `myfile`, the file with information about the intersections of sets 2 and 3 will be written to `myfile-2-3`.)

## Printing the Venn diagram

The Venn diagram displayed by Intersect can be printed to a file by choosing "Print Diagram" from the "Diagram" menu. The user is prompted to provide a filename.

*Note: The diagram is printed to the file in Postscript format. This format can be converted by several programs, including Adobe Illustrator, for use in other documents. The Postscript file can also be sent to a printer.*

## Saving the session

The current settings for the sets and display can be saved to a file by choosing "Save Session" from the "File" menu (alternatively by typing Ctrl-S). The format of this file is plain text, and is designed to be easily readable by a user. This allows users to easily construct Intersect session files directly with a text editor, or perhaps automatically via another program. See Appendix A for the format of the session file.

## Opening a saved session

Saved session files can be opened by choosing "Open" from the "File" menu (alternatively by typing Ctrl-O). The file information will be read from the session file, creating the appropriate controls for each set and filling in the information.

# Appendix A: Session file format

The format of the session file is simple, and designed so that users can create them using just a text editor. The format is as follows:

```
[Comments]
<blank line>

Set 1  [name]  color
file1
file2
…
file3

Set 2  [name]  color
file1
…
```

The parameters in brackets are optional. The sets are separated from each other (and the display line) by at least one blank line. Any number of blank lines can separate them, as long as there's at least one. The first line of the file should be blank if no comments are provided.

Here's an example of a saved session file:

```
C:/Desktop/Intersect/temp - Fri Nov 23 16:12:44 2001

Set 1  set one name  #0000aa
C:/Desktop/Intersect/w1.bst
C:/Desktop/Intersect/n10.bst
C:/Desktop/Intersect/n12.bst

Set 2  set two name  #aaaa00
C:/Desktop/Intersect/w7.bst
C:/Desktop/Intersect/w8.bst

Set 3  set three name  #00aa00
C:/Desktop/Intersect/w18.bst
```

© 2001 University of California

# Appendix B: Known bugs

The following are known bugs in the current version of Intersect. New bugs should be reported to spegg@mako.cgl.ucsf.edu.

© 2001 University of California

# Appendix C: Frequently Asked Questions

Frequently Asked Questions about Intersect version 0.1a:

1. *Why aren't the Venn diagrams drawn to scale?*

The Venn diagrams are not drawn to scale for a number of reasons. First, doing so is technically difficult, given that some of the intersections may be empty, requiring the positioning and sizing of the sets to change accordingly. While not impossible, it would require a lot more programming and probably run much too slowly to be of practical use.

Second, having a consistent drawing format makes repeated use easier. Locating specific intersections is quicker, and displaying empty ones as black often leads to the recognition that a certain set intersection is empty (as opposed to having the user hunt around for the particular intersection to see if it's displayed).

2. *How many sets of sequences can I use? Why don't I get a Venn diagram for five or more sets?*

There's no limit on the number of sets of files you can use with Intersect. Keep in mind, however, that the more sets you use, the more complex the output will become. Everyone has a personal limit to the number of intersections they're willing to look at.

The reason the limit for the Venn diagram is four is that it's not possible to construct a Venn diagram with all possible intersections represented using proper rectangles and more than four sets. The theoretical limit of the number of sets for which one can construct a complete Venn diagram using ovals appears to be 5. It can be done symmetrically, and looks pretty, but contains so many regions (31) that it becomes pretty unwieldy. Interestingly, a Venn diagram with 6 sets can be constructed using triangles, with even more intersections (63), but who really wants to deal with that?

3. *Why can't I print directly to a printer?*

In theory this is possible, but requires making some platform-specific code to deal with the differences in print spooling between various operating systems. This may be implemented in the future, but is not a high priority. It's only a few mouse clicks to print the current postscript file, and you should probably preview it anyway.

4. *Can I mix-and-match my database output file types?*

Yes, you can. Keep in mind, however, that the "probability" score reported in the intersection data may not have quite the same meaning in each file type. Currently supported file types are: BLAST(NCBI), WU-BLAST, FASTA, and PSI-BLAST.

5. *Can I use PSI-BLAST output files?*

Yes, you can. When the "E-value" score is reported from a PSI-BLAST file, it also contains the latest round (in parentheses) in which the given sequence was found.

6. *Who wrote this piece of *@%$!*

Intersect is the offshoot of research conducted in the laboratory of Patricia Babbitt at the University of California, San Francisco. It was written by Scott Pegg, Ph.D., with many constructive ideas from Walter Novak. If you actually like the program, or have some constructive suggestions, contact spegg@mako.cgl.ucsf.edu.

© 2001 University of California