SeqPup, version 0.4 development release, July 1995

Abstract

SeqPup is a biological sequence editor and analysis program usable on the common computer systems including Macintosh, MS-Windows and X-Windows. It includes links to network services and external analysis programs.

Features include

- multiple sequence alignment editor
- single sequence editor
- read and write several sequence file formats
- easy hand alignment features including colored bases and sliding
- automatic multiple sequence alignment thru ClustalW app
- automatic gel fragment alignment to contigs thru CAP app
- consensus, reverse, complement, degap operations
- restriction maps
- translate dna to/from protein using various codon tables
- find strings and ORFs
- automatic preference saving
- internet send mail
- internet sequence analysis services by email
- user-definable links to external analysis programs

NOTICE: This release is still unfinish, and has bugs. Please careful of trusting important work to it. Yet it may be useful to some of you as is.

SeqPup is being written by Don Gilbert using DCLAP, a free, portable C++ class application framework, and founded on the NCBI Toolkit, especially it's Vibrant user-interface section written primarily by Jonathan Kans. SeqApp/SeqPup was started in 1990 as sequence editor/analysis platform on which analysis programs from other authors could be easily incorporated into a useable interface. It was originally written with Apple Computer's MacApp application framework.

You can obtain this release thru anonymous ftp, gopher or http to iubio.bio.indiana.edu, in folder /molbio/seqpup. Versions are available for Macintosh, MS Windows, and various Unix/XWindows systems. The Internet locators to this software are

<ftp://iubio.bio.indiana.edu/molbio/seqpup/>
<gopher://iubio.bio.indiana.edu/11/IUBio-Software+Data/molbio/seqpup/>
<http://iubio.bio.indiana.edu/1/IUBio-Software%2bData/molbio/seqpup/>

Source code for this software is at <ftp://iubio.bio.indiana.edu/util/dclap/source/>.
Comments, bug reports and suggestions for new features (see below) are very welcome and should be sent via e-mail to <mailto://SeqPup@Bio.Indiana.Edu/>.

July 95: Version 0.4 of SeqPup. This release includes most of the features of its parent, SeqApp, as well as new features and corrections. Alignment window: shift & slide sequences, copy/cut/paste/undo sequence entries among windows; Restriction maps and pretty print output; useable child apps for mac, mswin, and unix.

SeqPup

version 0.4 development release
July 1995

¶ SeqPup Help

SeqPup is a biological sequence editor and analysis program usable on the common computer systems including Macintosh, Motif/X-Windows and MS-Windows. It includes links to network services and external analysis programs.

This program has already gone thru several changes since its start in September 1990. I don't expect it to mature for another year or two, as my prime programming time is holidays and weekends.

Comments, bug reports and suggestions for new features (see below) are very welcome and should be sent via e-mail to
<mailto://SeqPup@Bio.Indiana.Edu/>

With any bug reports, I would appreciate as much detail as is reasonable without putting you off from making the report. If you don’t have time to send detailed descriptions of problems, please do send comments and reports, even if all you say is "Good" or "Bad" or "Ugly".

Please include mention of computer hardware, and operating system software, including version. Describe how the problem may be repeated, if it is repeatable. If it is sporadic or only seen once, please also describe actions leading up to it. Include copies of data if relevant.

If you need to use land mail, send to

Don Gilbert
Biocomputing Office, Biology Department
Indiana University, Bloomington, IN 47405

¶ Fetching

You can obtain this software via Internet, using anonymous ftp, gopher or http to the IUBio server at iubio.bio.indiana.edu. It is located in folder /molbio/seqpup. Versions are available for Macintosh, MS Windows, and various XWindows/Unix systems. Please check the Readme files at this archive for recent news. Remember to use binary FTP to fetch the .zip and .gz binary files.

Internet resource locators for this software are

<ftp://iubio.bio.indiana.edu/molbio/seqpup/>
<gopher://iubio.bio.indiana.edu/11/IUBio-Software+Data/molbio/seqpup/>
<http://iubio.bio.indiana.edu/1/IUBio-Software%2bData/molbio/seqpup/>

Source code for this software is at

<ftp://iubio.bio.indiana.edu/util/dclap/source/>

You will need to fetch one of the program archive files for your computer system, its associated child app archive, and fetch the essential and optional items from the "all-systems" folder. For example, this would be

all-systems/ SeqPup.help, SeqPup.prefs, tables/*, seqs/*

plus

mac/ seqpup-mac-68k.hqx and seqpup-mac-apps.hqx

or

mswin-i86/ seqpup16.zip and spapp1.zip

or

unix/sun-sunos4-sparc/ SeqPup-sunos-mostat.gz and seqpup-apps.tar.gz

The current software distribution comprises the following items.

all-systems/ mac/ mswin-i86/ unix/

all-systems/	
SeqPup.help	- help file (RTF format) **essential**
SeqPup-help.text	- help file (plain text) *optional*
SeqPup.prefs	- preferences file (plain text) **essential**

tables/ - data files used by SeqPup, **essential**

codon.table dro.cod hum.cod renzyme.table

color.table eco.cod rat.cod tob.cod

appsrc/ - source to applications called by SeqPup, *optional*

ChildApp.c captest.seq fastDNAml.doc

cap.src/ clustalw.doc fastDNAml.infile

cap2.doc clustalw.src/ fastdnaml.src/

seqs/ - test sequence files, *optional*

23+28SrRNA.gb captest.fasta fastdnaml.phylip

5srna.gb dros.ig testre.map6

blue.seq ecolac.seq testreseq.gcg

mac:	- Macintosh, files are in binhex format
Readme	
seqpup-mac-68k.hqx	- SeqPup for Mac with Motorola 68000 processor
seqpup-mac-ppc.hqx	- SeqPup for Mac with PowerPC processor
seqpup-mac-apps.hqx	- child apps for mac, both 68k and PPC (fat binaries)

mswin-i86:	- MS Windows, files are in ZIP archive binary format
Readme	
seqpup16.zip	- Seqpup for MS Windows, 16-bit code
seqpup32.zip	- SeqPup for MS Windows, 32-bit code
spapp1.zip	- child apps for for mswin

unix:	- Unix, files are in TAR, Gnu ZIP format
dec-alpha-osf/	- DEC Alpha computer with OSF/1 Unix
sun-sol2-i86/	- Sun Solaris 2 on Intel 80x86 processor
sun-sunos4-sparc/	- Sun SunOS4 on SPARC processor (or Sol2)
sun-sol2-sparc/	- Sun Solaris 2 on SPARC processor
sgi-irix/	- Silicon Graphics Iris

unix/dec-alpha-osf:
Readme SeqPup.gz seqpup-apps.tar.gz

unix/sgi-irix:
Readme SeqPup.gz seqpup-apps.tar.gz

unix/sun-sol2-i86:
Readme SeqPup.Z seqpup-apps.tar.Z

unix/sun-sol2-sparc:
Readme SeqPup.gz seqpup-apps.tar.gz

unix/sun-sunos4-sparc:
Readme
SeqPup-sunos-mostat.gz - Motif libraries are included (will run on SunOS 4 or Solaris 2 lacking Motif libraries)
SeqPup-sunos-dyn.gz - Motif libraries are not included
seqpup-apps.tar.gz - child apps for SPARC

¶ Installing

SeqPup is distributed over the Internet in archive files. The archive format used is commonly available on the computer system you use (HQX self-extracting for Macintosh, ZIP for MS Windows, and tar + Gnu ZIP for Unix). There is one primary program, several document and data files, examples, and child application programs.

The current organization of files used by the program is:
SeqPup -- execuable, called "seqpup.exe" in MSDOS
SeqPup.help -- this document, in Microsoft RTF format
SeqPup.prefs -- settings for the program, in text format.

tables/ -- data files, required for Restriction maps, translate and some other functions. These are standard bioinformatics data files available and updateable from various sources.
codon.table -- table of codon preferences, in GCG format
renzyme.table -- REBase data file, for restriction maps, in GCG format
color.table -- table of color values for display of bases
hum.cod, tob.cod, eco.cod, and other codon preference tables that can substitute for codon.table at your preference

apps/ -- a selection of external analysis applications.
clustalw -- multiple sequence alignment
cap2 -- contig alignment
fastDNAml -- phylogenetic analysis of sequences

Note for Sun systems: This program requires theMotif run-time libraries that are commonly found on other XWindow systems. Motif is not standard on SunOS and is not part of Solaris until verson 2.4. If you have Solaris 2.3 or earlier, or SunOS, and do not know that your system includes Motif, then you will need the version with statically bound Motif libraries (SeqPup-sunos-mostat.gz).

If you have Solaris 2.4, or a version where Motif libraries are present, you may still need to configure the system to let SeqPup know where they are. In a Solaris 2.4 system, where motif lives in /usr/dt/lib by default, this may be needed to run successfully:

```
   setenv LD_LIBRARY_PATH   "$LD_LIBRARY_PATH":/usr/dt/lib
```

¶¶ Installing preferences

In addition to these two folders and three SeqPup files, the program will automatically create a personal preferences file in you computer when you first run it. These preferences come from the SeqPup.prefs file. The preferences file created on your system will be something like this

System Folder:Preferences:seqpup.cnf	- MacOS
c:\windows\seqpup.ini	- MS Windows
~/.seqpuprc	- Unix

The program will save various configuration information to this file. You may edit this with a text editor. You may delete it and a new one will be generated from the SeqPup.prefs file. You may not edit it while the program is active (any such changes are lost). When the program is updated in the future, new preferences are added, using the label

```
[version=123]
```

to indicate the version number.

The preference file format is as follows:

- Logical sections are indicated in brackets [section].
- Variables are denoted with a "name=value" format.
- Line starting with ";" indicates a comment and will be ignored.

The current release of the program may require some fiddling to install correctly. This is a known problem, and will be corrected in future releases. You will want to look at and probably edit the file "SeqPup.prefs".

The following sections are important in getting the program to work right, and may need to be edited.

```
[paths]
temp=
tables=tables
apps=apps

[data]
codon=tables:codon.table
renzyme=tables:renzyme.table
color=tables:color.table
```

If you use this on a Unix system or an MS DOS system, the current configuration should work if you start the program from its folder, e.g.,

```
 cd /path/to/seqpup/
./SeqPup
```

But as is common on Unix, if you want to install this for use from any directory, you will currently need to edit the prefs file and put a fixed path to the SeqPup folders in it, as

```
[paths]
temp=/tmp
tables=/long/path/to/seqpup/tables
apps=/long/path/to/seqpup/apps
```

If you run SeqPup first, then decide to change parts of the prefs file, you can have all users prefs be updated if you add the new prefs after a new version number. This is the procedure:
a) add a higher version number at the end of the SeqPup.prefs file

```
[version=6]
```

b) add changed preference sections and values after that. You need not remove or edit the original values (I hope...).

So for instance if the highest verson value in the prefs file is 5, then add this at the end of the SeqPup prefs to get all users preferences updated:

```
[version=6]

[paths]
tables=/new/path/to/seqpup/tables
apps=/new/path/to/seqpup/apps
```

An important caveat with this: New distributions of SeqPup will use new version values to trigger preference updates. If the new distribution has a lower version value than you have used, it won't trigger an update.

Child applications are configured for use with the SeqPup.prefs file. Please see below the section **Child Tasks.**

¶ Source code and DCLAP

SeqPup is built on an object-oriented application framework, written in C++, called DCLAP. This framework is designed to speed the development of easy to use, complex programs with a rich user-interface. At this point, DCLAP is still an unfinished framework, lacking in documentation. However, it is rich enough at this point to build complex programs like SeqPup.

DCLAP includes the following segments

DClap/	-- basic application framework, including command, control, dialog, file, icon, list, menu, display panel, table view, mouse tracker, child application, window and view classes.
Drtf/	-- rich text display handlers, including RTF, HTML document, PICT and GIF image format readers.
DNet/	-- Internet connection tools, including TCP/IP, SMTP, Gopher and preliminary HTTP classes.
DBio/	-- Biocomputing methods, included biosequence, restrict enzyme, sequence editor, seq. manipulator, seq. output classes.

New applications can be built to employ and reuse these classes fairly quickly. Variations on the current methods are simple to add in the class derivation method of C++. For instance, new document formats can be added on the Drtf display objects, and new sequence manipulations can be added in the biosequence handlers, by building on current methods.

DCLAP rests upon the NCBI toolkit, including the Vibrant GUI toolkit, which is designed for cross-platform functioning. The successful genome data browser Entrez is written with the NCBI toolkit.

All of this source is available without charge for non-profit use (see copyright below). The NCBI toolkit portion is further available for profit use, and such arrangements may be made for use of DCLAP.

DCLAP will never compete with commercial programming frameworks, but it has the virtue of being freely available and redistributable, and includes support specifically for biocomputing applications. If you are undertaking a biocomputing project requiring a rich user interface, and wish it to run on multiple computer platforms, this may be a worthwhile choice, especially if you wish to redistribute your source code for the benefit of the scientific community.

The DCLAP developer archive is at <ftp://iubio.bio.indiana.edu/util/dclap/>
Please contact Don Gilbert for further information on using this framework in other applications.

¶ Copyright

gilbertd@bio.indiana.edu
Biology Dept., Indiana University, Bloomington, IN 47405

You may use this program for your personal use, to provide a non-profit service to others.
You may not use this program in a commercial product, nor to provide commercial service, nor may you sell this code without express written permission of the author.
You may redistribute this program freely. If you wish to redistribute it as part of a commercial collection or venture, you need to contact the author for permission.

The source code to this program is likewise copyrighted, and may be used, modified and redistributed in a free manner. Commercial uses of it need prior permission of the author.

Any external applications that may distributed with SeqPup are copyrighted by their respective authors and subject to distribution provisions as described by those authors. At present this includes ClustalW, by Des Higgins, CAP2 by Xiaoqiu Huang, and FastDNAml, written by Joseph Felsenstein with modifications by Gary Olsen, Hideo Matsuda and Ross Overbeek, is copyrighted by University of Washington and
Joseph Felsenstein.

Distribution of external analysis applications with this program is done as a convenience for users, and in no way modifies the original copyright. If there is a problem with this, instructions to users for obtaining and installing external applications will be substituted.

No warranty, express or implied, is provided with this software. The author is trying to produce a good quality program, and will incorporate corrections to problems reported by users of it.

¶ Views

There are four main types of views or displays in SeqPup:

A multiple-sequence view which is the primary display when you open a sequence document; the single sequence editting view; various print views which result from an analysis, like the Restriction map; and dialog views where you control some function.

Many of these views have dialog controls -- push buttons, check boxes, radio controls and edittable text items -- to let you fine-tune a view to fit your preference. Many of these views also will remember your last preferences.

When a view has editable text items, including the sequence entry views, most usual undo/cut/copy/paste features will work.

Two or more views of the same data are possible. Some of these are truly views of the same data -- changes made in one view are reflected in another. Other views are static pictures taken of the data at the time the analysis was performed -- later changes to the data do not affect that picture.

¶¶ Aligned multi-sequence view

The main view into a sequence document is the multiple sequence editor window, which lists sequence names to the left and sequence bases as one line that can be scrolled thru. Bases can be colored (now only nucleic colorings) or black. Sequence can be editted here, especially to align them, and subranges and subgroupings can be selected for further operations or analysis. Entire sequence(s) can be cut/copied/pasted by selecting the left name(s). Mouse-down selects one. Shift-mouse down selects many in group, Command-mouse down selects many unconnected. Double click name to open single sequence view. Select name, then grab and move up or down to relocate.

Select the lock/unlock button at the view top to lock/unlock text editting in the sequence line. With lock on (no editting) you can use shift and command mouse to select a subrange of sequences to operate on.

Bases can be slid to left and right, like beads on an abacus, when the edit lock is On (now default). Select a base or group of bases (over one or several sequences), using mouse, shift+mouse, option+mouse, command+mouse. Then grab selected bases with mouse (mouse up, then mouse down on selection), and slide to left or right. Indels "-" or spacing on ends "." will be added and squeezed out as needed to slide the bases. See also the "Degap" menu selection to remove all gaps thus entered from a sequence.

¶¶ Single sequence view

For entering/editting a single sequence, this view displays one sequence with more info and control. Edit the name here (later other documentation). Bring out this view by double-clicking sequence name in align view, or choosing Edit from Sequence menu.

¶¶ Print views

Various analyses provide non-editable displays. These are usually save-able as PICT format for editting in your favorite MacDraw program, or print-able.

¶ Data files

SeqPup uses plain TEXT type files for its primary sequence data. These files can be exchanged without modification with many other sequence analysis programs. SeqPup automatically determines the sequence format of a data file when openning it. You have an choice of several formats to save it as. As of this writing, the GenBank format is prefered (see bugs).

The program looks in the folder "tables" for text files containing various data. At present these files include "codon.table", "renzyme.table" and "color.table".

There is a "SeqPup.prefs" file which stores various user options like window positions, mail address, child tasks. This is described more in the Install and Child Apps sections.

Various temporary files are created for child tasks, generally in the :Apps: folder. Currently you cannot run the Child Tasks portion of SeqPup from a locked file server because these temporary files need to be created where the child applications reside. Otherwise, SeqPup should operate from a locked fileserver properly, and can be launched by several users at once.

¶¶ Restriction Enzyme Table

The file called "renzyme.table" contains restriction enzyme data, as distributed in REBASE by R.Roberts. The format used is identical to that used by GCG software.

```
            { documentation ...}

Commercial sources of restriction enzymes are abbreviated as follows:

                    A        Amersham (12/91)
                    B        BRL (6/91)
                    ...
                    X        New York Biolabs (4/91)
                    Y        P.C. Bio (9/91)

..   {< separates data}
;AatI      3 AGG'CCT      0 !   Eco147I,StuI                       >OU
AatII      5 G_ACGT'C    -4 !                                      >EJLMNOPRSUVX
AccI        2 GT'mk_AC     2 !                                     >ABDEIJKLMNOPQRSUVXY
;AccII     2 CG'CG        0 !  Bsp50I,BstUI,MvnI,ThaI              >DEJKQVXY
;AccIII    1 T'CCGG_A      4 !   BseAI,BsiMI,Bsp13I,BspEI,Kpn2I,MroI   >DEJKQRVY
;Acc65I     1 G'GTAC_C      4 !   Asp718I,KpnI                      >DFNY
```

¶¶ Codon Table

The file called "codon.table" in folder "Tables" is used for translation of nucleic to protein sequence, and for backtranslation. This file may be replaced with a table of your choice in the following format (this format is identical to that used by GCG software codon tables).

```
            { any documentation... }

AmAcid   Codon     Number     /1000     Fraction    ..    {< data separator}
Gly      GGG      1743.00      9.38      0.13
Gly      GGA      1290.00      6.94      0.09
...   { continue for 64 codons }
```

¶ Features

The following topics describe main features found in the SeqPup menus.

¶¶ File

New will create an align view of sequence data. New Text will create a plain text document, which is the format of the sequence data files also.

Open will open an exising file. The default choice will open a file of sequences into a new window. You can choose "Sequence, append", or hold down the SHIFT key, to open a sequence file and append it to an existing alignment window.

Other **Open** options include opening a plain text file, a file of phylogeny trees in Newick format (see Phylip documentation), or a Gopher document.

Save, Save as, Save a copy in, all will save the current document to disk files. Revert will restore the open align view to the last version saved to disk.

Save selection, Saves only highlighted sequences to a new disk file. Doesn't affect save status of current full alignment document.

Print setup, print will print the current view.

Help brings up a view to page thru the help file.

Preferences will set some user preferences.

¶¶ Editing

Undo, **cut**, **copy**, **paste**, **clear**, **select all** -- these standard mac commands will operate on text as well as on sequences in (hopefully) intuitive, usual ways.

Find, Find same, Find "selection" will search for strings in text.

Replace, replace same will replace target strings (not yet enabled).

¶¶ Sequence manipulations

New sequence -- append a new, blank sequence to the sequence document.

Edit -- open single sequence editting view for selected items.

Reverse, **Complement**, **Rev-complement** -- Reverse, complement or reverse+complement a sequence. Works on one or more sequences, and the selected subrange.

Rna-Dna,Dna-Rna -- Convert dna to rna (t->u) and vice versa. Works on one or more sequences, and the selected subrange.

Degap -- remove alignment gaps "~". Works on one or more sequences, and the selected subrange. Gaps of "-" are locked and not affected by Degap. Works on one or more sequences, and the selected subrange.

Lock Indel & Unlock Indel -- Convert from unlocked gaps "~", to locked gaps "-". Unlocked gaps will disappear and appear as needed as you slide bases left and right. Locked gaps are not affected by sliding nor by Degap. Works on one or more sequences, and the selected subrange.

Consensus -- generate a consensus sequence of the selected sequences.

Translate -- translate to/from amino acid. Relies on Codon.Table data.

Pretty print -- a prettier view of a single or aligned sequences. Use these views to print your sequences. Printing from the editing display will not be supported fully, and may not print all of your sequence(s).

Restriction map -- Restriction enzyme cut points of selected sequence. Also protein translation options.

Dotty plot -- provide a dot plot comparison of two sequences.

Nucleic, amino codes -- These provide both reminders of the base codes, and a way to select colors to assocate with each code (new in v 1.9a). See below for some discussion of the two "aa-color" documents that now ship with SeqPup.

¶ Child Tasks

The "ChildApps" menu lets you link SeqPup with external sequence analysis programs that you or others may write. SeqPup can be configured to launch any other application, and to send it sequence data and command information. When the child program is finished with its analysis, SeqPup can open and display results files from the child in a variety of formats, including text, biosequence, PICT, RTF and GIF. On Macs, the ChildApps menu requires System 7 to operate.

The general design of child applications is taken to be data analysis programs that have a simple command-line user-interface, and that take input data from a file or from the system "standard input" file (stdin), and that write outputs to files and to two system standard files "standard output" (stdout) and "standard error" (stderr). This is how many existing analyses programs work, and it is very straightforward to program this basic kind of user-interface.

The value of SeqPup joined with these kinds of programs is that the SeqPup can concentrate on providing an easy-to-use interface for biologists, and the analysis application can concentrate on data analyses, without having to add a lot of software baggage to provide a more usable interface.

A desired addition to SeqPup will be a dialog to configure new and current child tasks. However, at present this needs to be done by using a text editor to change the SeqPup.prefs file.

¶¶ Configuring child applications

You can add new child apps by editing the text file SeqPup.prefs. You will need to update the section [apps] with a new line for you new app, then install a new section, [newappname]. You will also need to increase the [version=#] value, as described above in the **Installation** section, for the program to take notice of your changes.

The [apps] section contains a list of child app sections, and the menu title string. E.g.,

```
[apps]
clustal=ClustalW Multiple align...
```

The clustal= line says there is a child app section called clustal, and its menu title is "ClustalW Multipla align..."

Then the section for [clustal] includes these variables
desc= descriptive string, displayed in the launch dialog
path= path to application, using variables defined in [paths] section
help= path to help document, ditto
cmd= command line passed to application
infile= path/name of input data file, using variables defined in [paths] section
seqformat= format for sequence input data file
minseq= minimum number of sequences required for application
outfile1= first output file, and file format in pseudo-mime notation
outfile2= second output file, and file format in pseudo-mime notation
... etc... for more output files.

All the lines which specify file paths should use the variables defined in the [path] section for an easy way to make these descriptions portable to other systems. The [paths] section specifies variables for file paths then gives their complete specification on the local file system, e.g.,

```
[paths]
```

```
    temp=/tmp
    apps=/long/path/to/seqpup/apps
```

Then in an application variable use the syntax "$pathvar:" to insert the local path variable. For example, use

```
    help=$apps:clustalw.doc
```

This will be translated by the program to

```
    help=/long/path/to/seqpup/apps/clustalw.doc
```

If no path is specified, the default path will generally be, on Macintosh, where the program file was when launched, and on Unix and MSDOS, where the command line was executed from.

The command line variable "cmd" should specify files and other parameters that the child application needs to read.

The current selection of "seqformat" sequence input formats includes the following:
genbank, fasta, embl, nbrf, pir/codata, gcg, msf, phylip, paup/nexus, asn1.

The current selection of pseudo-mime notations known by SeqPup used to specify the return data formats includes biosequence formats, basic text and image formats:
biosequence/genbank, biosequence/fasta, etc.. for sequence formats
biotree/newick - newick style phylogenetic tree, not yet displayable
text/plain, text/rtf, text/html - text file formats
image/pict, image/gif - image file formats

Seqformat for the input file now is not in pseudo-mime format, but may change to that for consistency with output formats. That would be "biosequence/fasta" instead of just "fasta".

¶¶ Current child app configuration

Current configuration of child apps in SeqPup.prefs file

```
    [apps]
    clustal=ClustalW Multiple align...
    cap=CAP Contig assembly...
    fastdnaml=FastDNAml...
    ;dnaml=Phylip DNAML...

    [clustal]
    desc=A multiple sequence alignment application
    path=$apps:clustalw
    help=$apps:clustalw.doc
    ;cmd=/infile=$temp:clustal.inseq /outfile=$temp:clustal.outseq /output=gcg /align
    cmd=/infile=clustal.inseq /outfile=clustal.outseq /output=gcg /align
    infile=$apps:clustal.inseq
    seqformat=nbrf
    minseq=2
    outfile1=$apps:clustal.outseq    biosequence/msf
    outfile2=$apps:clustal.dnd biotree/newick

    [cap]
    desc=A gel contig assembly application
    path=$apps:cap2
    help=$apps:cap2.doc
    ;cmd=$temp:cap.inseq $temp:cap.outseq 20 85
    cmd=cap.inseq cap.outseq 20 85
```

```
infile=$apps:cap.inseq
seqformat=fasta
minseq=2
outfile1=$apps:cap.outseq  biosequence/fasta
dlog1=input name=minoverlap value="20" title="Min. number bases overlap"
dlog2=input name=pctmatch value="85" title="Percent match"

[fastdnaml]
desc=A faster DNA max likelihood phylogeny estimator
path=$apps:fastDNAml
help=$apps:fastDNAml.doc
cmd=
infile=$apps:infile
seqformat=phylip
minseq=2
outfile1=$apps:outfile     text/plain
outfile2=$apps:checkpoint.0      text/plain
```

¶ Internet

The Internet features of SeqPup let you interchange ideas and data with people and biocomputing services around the world. If your Mac is connected already to the Internet, you probably are familiar with electronic mail and some of its uses.

SeqPup includes a selection of network access features in the developing area of networked biocomputing. You will find access to me, at least to get comments and bug reports to me, very easy. There is a feature to send and receive e-mail, as well as mail links to customized e-mail services. These include searching for sequence similarity via BLAST and FastA programs on the Genbank/Intelligenetics computers, fetching sequences, data and software from Genbank and EMBL.

There is now an feature called Gopher, which gives you access to a wide range of information services now developing on the Internet. Gopher is something like Telnet or FTP (file transfer), but also different. It includes some of the keyword searching features of WAIS (Wide Area Information Services). There are currently several biology gopher services found around the globe. These include fast and up-to-date keyword searches of GenBank, EMBL, PIR and other important biology databanks.

¶¶ Internet requirements

All features of this menu depend on a network link to the Internet, and
Mac: MacTCP software from Apple Computer, or equivalent.
MS Windows: WinSock.dll software from various vendors
Unix: TCP should be standard software

If you have problems in general with SeqPup network functions, make sure that other TCP-based applications work on your computer before reporting the problem. You may need to work with computer support people at your site to iron out general network problems.

¶¶ Mail Preferences

The mail prefs dialog asks for your return e-mail address, and your preferred SMTP mail host. These addresses may be similar.

Return e-mail address: This is where another person should send mail so it will reach you.
Example: Bob.Jones@Bio.Indiana.Edu
or: bjones@sunflower.bio.indiana.edu

SMTP Mail host: This is the internet address of the computer thru which SeqPup will send out mail to the rest of the world.
Example: Sunflower.Bio.Indiana.Edu

¶¶ Send Mail

Send an electronic mail message. You must enter an address to send to, and have entered your return address in the mail preferences dialog.

¶ Mail-based Search and Fetch

Various network resources provide biocomputing services thru e-mail. These include retreiving sequence entries from the various databanks (GenBank, EMBL, PIR), fetching help documents, and searching for sequences in the databanks that match your query sequences.

¶¶ Sequence Searching

Mail based servers for searching databanks against your query sequence include FastA and BLAST searches for nucleic or protein sequences at GenBank/IntelliGenetics, and protein searches at PIR.

¶¶ Gene Prediction

There are, as of Feb 1992, two e-mail based services for analyzing nucleic acid sequences and predicting gene structure. These services use a variety of analyses and combine them to provide their best "guess" at gene structure.

Geneid is an Artificial Intelligence system for analyzing vertebrate genomic DNA and prediction of exons and gene structure (1). A prototype is implemented as a fast, automatic email-response system.

Grail is an interface to a system which will ultimately provide automated gene assembly from DNA sequence data. Currently the system provides analysis of protein coding potential of a DNA sequence. The coding recognition module (CRM) uses a multiple- sensor neural network approach to identify coding exons than are at least 100 bases long.

Both of these services ask that you register once before using them.

¶¶ Sequence Fetching

Mail based servers for fetching databank entries include services from GenBank/NCBI, Univ. of Houston, PIR, and EMBL.

¶ Color Selections

You can create your own color selection for alignment display by choosing the Nucleic codes or Amino codes dialogs from the Sequence menu. These dialogs provide color "buttons" for each base. Click a button to get a color picker dialog where you can change the currently assigned color. Your selection can be saved to disk file as an amino color or a nucleic color document. You can reload such a color scheme by clicking open this document, or by choosing it from the File "Open…" dialog.

A few early users of this new version provided two of the color amino selections that ship with SeqPup. Here are their descriptions.

Date: Fri, 28 May 1993 20:07:26 -0500
From: ahouse@hydra.rose.brandeis.edu (Jeremy John Ahouse)
Subject: implemented aa colors for pre-rel seqApp

```
Don Gilbert (& Phil Carl),
        I have implemented Phil Carl's(*) modest proposal.
Some of the suggestions were not possible, so I made changes.

        Jeremy Ahouse

Phil's suggestion is interspersed with my additions:

Well, I have (as they say) a modest suggestion.  I suppose what people
are really seeking are 20 colors for 20 amino acids.  I have a preliminary
proposal based on classifying the amino acids into chemical groups and
finding what seems to me to be easy pneumonics for each group.  Thus
I would propose:


Red for acidic amino acids; Glu, Asp
(since red is a common danger signal and acids are dangerous
(well maybe not amino acids, but it's a start))
hue:        65500
saturation: 65000
brightness: 50000

Blue for basic amino acids; Lys, Arg, His
(blue and basic both start with "b")
hue:        44000
saturation: 65000
brightness: 50000

White for hydroxyl amino acids; Ser, Thr (as in whitewater)
(this was not possible so I chose a cool "whitewater" color)
hue:        33000
saturation: 65000
brightness: 50000

Green for amide amino acids; Asn and Gln
(since glutamine and asparagine rhyme with green)
hue:        22000
saturation: 65000
brightness: 50000

Yellow for sulphur amino acids; Cys, Met
(this one's obvious)
hue:        12000
```

saturation: 65000
brightness: 60000

Black for hydrophobic amino acids; Ala, Val, Leu, Ile
(Black is the opposite of white and so if white is for hydrophilic hydroxyl amino acids black is a natural for hydrophobic ones)
hue: 00000
saturation: 00000
brightness: 00000

Orange for aromatic amino acids; Tyr, Phe, Trp
(since "orange" sounds a little like "aromatic" and oranges are aromatic (if that suits you better))
hue: 7000
saturation: 65000
brightness: 60000

Purple for proline; Pro
(since both have "prl" in them)
hue: 51000
saturation: 65000
brightness: 60000

Grey for glycine; Gly
(since both start with "g" and grey is sort of blah-like glycine)
hue: 00000
saturation: 00000
brightness: 30000

*Phil Carl
Assoc. Director
Program in Molecular Biology and Biotechnology
University of North Carolina, Chapel Hill

=======================

Date: Mon, 7 Jun 1993 15:50:09 +0200
From: Heikki.Lehvaslaiho@Helsinki.FI (Heikki Lehvaslaiho)
Subject: aa colors

Hi,

I am including a file with amino acid color codes that are used in Steven Smith's GDE. This scheme was not mentioned in the Usernet discussion, but I've grown accustomed to it. At least, it is no worse that any other of the myriad possible coloring choices.

If you haven't got other schemes in files yet, drop me a note and I'll see what I can do.

GDE aa-colors:

		24-bit			Mac		
COLOR	AA	R	G	B	R	G	B
Magenta	AGPST	255	000	255	65535	0	65535
Black	BDENQZ	000	000	000			
Red	C	225	000	000	57600	0	0
Blue	FWY	000	000	255	0	65535	65535
Light blue	HKR	000	192	192	0	49344	49344
Green	ILMV	000	192	000	0	49344	0
Gray	JOUX	145	145	145	37265	37265	37265

```
          -Heikki
```

¶ Bugs

v0.4 Known bugs and missing features:

General:

- Single sequence editor (Sequence/Edit) is very slow for long sequences (6,000bases)
- Repeated copy/cut/paste of the alignment window entries might cause problems. Copy of sequences between windows may lead to problems. Please let me know if you see this.
- copy/cut/paste/undo and clipboard functions may not be working as smoothly in as many contexts yet as they should be.
- Sequence menu items not yet ready : Dot plot.
- Sequence/translate when done on a subsequence selection, will now leave excess nucleic bases in selection?
- Edit menu items not ready yet: Show clipboard
- Find-again may still fail if there was no previous Find (fixed now in source)
- Internet menu needs testing & reworking - I haven't tested any of the e-mail services listed since last year.
- Nucleic codes picture shows PICT processing bug -- misplaced text, and an error in biology -- complement of W is W, not S, and complement of S is S, not W.
- Rich Text, PICT and GIF Image format displays all have various display glitches. Documents in these formats will be displayed for the most part but some RTF or images may show mistakes. Some of this is platform-dependent.
- The current release may require some fiddling to install correctly (see Installing).
- pretty print and map output cannot yet be saved in a picture file format, only in text file format.
- a dialog function to install new child apps, and to remove and edit current ones, is desirable. Configurable dialog functions for any child app are desirable.
- Windows menu should list current windows directly, but base toolkit doesn't yet handle menu item additions/deletions after program start.
- Pretty print display should allow user selections for alignment boxes and shading.
- Should File/Open> open sequence / open text,rtf,etc / open append submenu items be promoted to main File menu items for easier use?
- Replace-find function is not ready yet.
- Option dialogs for selecting tables (codon, renzyme, color) should allow direct editing of table values.
- Selection highlight in align window leaves messy box after deselection (cosmetic bug).
- Single edit window key checking beep works on name as well as sequence data.
- Scroll bar is slightly misplaced in rich text window.
- horizontal scroll in pretty print windows doesn't work right.
- item selection in rich text and pretty print windows doesn't display properly.
- documentation (this file) is not complete yet. Need more description and examples on how to use the methods.

MS Windows specific:

- Text editing in alignment window doesn't track properly when window is scrolled.
- standard system copy/paste/cut/undo command keys are not yet supported.
- application crashes or lockups may be too frequent to make this currently usable. Some feedback from users will help decide which bugs to attack first.
- printing has not yet been tested from MS Windows (no printer on my mswin box).
- select-all in align view highlights part of sequence lines when it should not.

- About-app image doesn't display -- due to draw pict bug with non-256 color images.
- Some dialog text items (e.g., seqprint prefs reenzyme range) don't show and allow edits.
- program chokes on quit or on window close sporadically.
- FIXED: Child app: SeqPup died after child app terminated -- problem w/ LaunchDialog delete..

XWindows specific:

- Text editing in the alignment window (unlock text) doesn't yet work -- the sequence disappears in edit mode.
- command keys are not yet supported as on Mac and MSWin systems.
- there is frequently an XText error message written to command line or console just before certain XWindow operations. This doesn't seem to cause problems yet, and can hopefully be ignored until I trace and squash it.
- Child Apps fail in various ways

  -- CAP seems most likely to succeed completely.

  -- ClustalW and FastDNAml may be launched and run properly, but SeqPup will fail to automatically open their results files. Possibly the fork/exec doesn't terminate properly.

  -- Launch dialog window caused some conflict w/ lauch of app such that when it is deleted, the main app dies. Currently this launch dialog is disabled in XWindows (menu item launches app directly).
- XWindows versions always crash/core dump when Quit is chosen. This is an annoyance but doesn't seem to impair use. I'll track it down.
- application can get confused at times about which window is active and front most. This is obvious when a function such as copy/paste acts in the wrong window. Sometimes repeated selection of items in front window will un-confuse the app.
- You will probably need to edit the SeqPup.prefs &/or ~/.seqpuprc file to put in fixed paths to tables/ and to apps/, and to change the paths to child app data files.
- There is no printing for X Window systems. This is not really my problem as much as it is the X Window design committee's problem. Among **over 25 pounds** of X Window programming books I have, **THERE IS NO MENTION OF HOW TO DEVELOP SOFTWARE FOR PRINTING FROM X WINDOWS APPLICATIONS.** I doubt the XWin committee views this as important, but most software users I know like to print documents on occasion. I'll handle this oversight sometime, but it won't be simple. Macintosh and MS Windows both provide methods for printing.
- SGI version only -- messes up with RTF and probably Text document display.
- About-app image doesn't display -- due to draw pict bug with non-256 color images.

¶ Future Features

Here is a list of things which may be added to SeqPup in the future, depending on your interest. Please send in your suggestions! What do you want to see to make this a good biosequence editor and analysis program?

Sequence documentation handling. Currently no provisions for documentation per sequence. This will at least change to a window for any comments and saving it into files (where file format permits). Possibly I will put effort into dealing with the features, references, etc., in a fashion along lines of Genbank/EMBL documentation structure &/or Authorin documentation. Your comments on the importance of this are desired.

Feature table parsing -- pull out subsequences from Gen/EMBL feature info.

Align, single sequence pretty print -- header, page numbering user prefs should be added.

Restriction map -- Could use some speed-up. Some would like graphic map (i.e., one line or circle w/ cut points per zyme).

Simple protein analysis routines, better protein handling.

Methods to transparently use networked child tasks (e.g., on fast compute servers).

¶ History

SeqApp was started Sept. 1990 as MacApp sequence editor/analysis platform on which analysis programs from other authors, typically command line w/ weak user interfaces, could be easily incorporated into a useable Mac interface.

July 95: Version 0.4 of SeqPup. This includes most of the features of its ancestor SeqApp. Alignment window: shift & slide sequences, copy/cut/paste/undo sequence entries among windows; Restriction maps and pretty print output; useable child apps for mac, mswin, and unix.

v0.4 corrections:

- File/Open for non-sequence data (text, rtf, etc.) has alternate open menu, to distinguish from sequence data. Added sequence append-open.
- Cut/copy/paste/undo for align-seq view now available
- Sequence menu items that are now ready: Consensus, Pretty print, Restriction Map, nucleic & amino codes. Some of these need further work (pretty, remap options).
- Child apps usage improved, may need more work though.
- The Mac/68K, Mac/PPC, MSWin, Unix now do Child applications.
- Include ClustalW, CAP, FastDNAml, child apps
- Restriction map function is extensively revised and improved.
- FindORF and Find string functions added
- Printing for pretty print, r.e.map now functional on Mac (and maybe MSWin)

v0.4 Known bugs and missing features (see above Bugs section for fuller list):

- Character editing (unlocked text) in the alignment (main) window is not working on Xwindow systems, and may be bugging in MSWindow and Mac systems.
- Single sequence editor (Sequence/Edit) is very slow for long sequences (6,000bases)
- Sequence menu items not yet ready : Dot plot.
- Child Apps fail in various ways on MSWindows and Unix systems.

-- CAP seems most likely to succeed completely.

-- ClustalW and FastDNAml may be launched and run properly, but SeqPup will fail to automatically open their results files.

- MSWindows and XWindows versions are less stable than Mac versions.
- XWindows versions reliable crash/core dump when Quit is chosen. This is an annoyance but doesn't seem to impair use.
- Internet menu needs testing & reworking - I haven't tested any of the e-mail services listed since last year.
- Nucleic codes picture shows PICT processing bug -- misplaced text, and an error in biology -- complement of W is W, not S, and complement of S is S, not W.
- Repeated copy/cut/paste of the alignment window entries might cause problems. Please let me know if you see this.
- There is no printing for X Window systems.

21 Mar 95: Second release of SeqPup, version 0.1. This release has more parts of the SeqApp program put into it. This includes some alignment view manipulations, limited use of child applications, some undo-able commands, choosing data tables for colors, codon and r.enzymes. This release also includes much of the basics of GopherPup, including display of RTF, HTML, PICT, GIF document formats. However there is still some work to be done to let you open these w/o interpreting them as sequence data.

This release has just a Mac PowerPPC (SeqPup/PPC) and Mac 68000 processor (SeqPup/68K) versions. When more of the basic bugs are worked out, I'll try Sun and MSWindows versions.

v0.1 Known bugs/missing features:

- Use of character editing (unlocked text) in the alignment (main) window will lead to a crash after a few windows have been opened/closed or other manipulations performed.
- File/Open for non-sequence data (text, rtf, etc.) may well mistakenly identify them as sequence data. File/New is probably not doing anything useful, or bombing.
- Single sequence editor (Sequence/Edit) is very slow for long sequences (6,000bases)
- Single seq. editor may be failing in various ways (I've not looked at it carefully yet).
- No cut/copy/paste/undo for align-seq view yet (coming soon I hope).
- Internet menu needs reworking - I haven't tested any of the e-mail services listed there since last year.
- Sequence menu items not yet ready : Consensus, Pretty print, Restriction Map, Dot plot, nucleic & amino codes.
- Child apps usage needs more development to work smoothly.
- The Mac/68K version fails when using Child applications.
- Only the ClustalW child app is ready for distribution (may have FastDNAml, CAP, and DNAml soon -- let me know of programs you would like to see here).

1 Mar 94: First public release of SeqPup, version -1.

It has plenty of bugs and missing features, including:

no Undo (this is a real bite to those used to it)
mostly no cut/copy/paste/clear
limited printing of documents or views
mostly no align-view manipulations (move,cut/copy,edit in place, shift, ...)
no pretty print views
no restriction maps
no dot plots
no ...
problems w/ window display & keeping track of active window (x,mswin)

I'll be adding back many of these features from the Macintosh SeqApp as time permits.

12+ June 93, version 1.9a157+ -- a semi-major update, and time extension release with various enhancements and corrections. These include

-- lock/unlock indels (alignment gaps). Useful when sliding bases around during hand alignment, to keep alignment fixed in some sections.
-- color amino (and nucleic) acids of your choice.
-- added support for more sequence file formats: MSF, PAUP, PIR. SeqApp now relies on the current Readseq code for sequence reading & writing.
-- save selection option to save subset of bases to file.
-- addition the useful contig assembly program CAP, written by Xiaoqiu Huang.
-- major revision of preference saving method (less buggy, I hope)
-- major revision of the underlying application framework, due to moving from MacApp 2 to MacApp 3.
-- fixed a bug that caused loss of data when alignment with a selection was saved to disk.

5 Oct 92, version 1.8a152+ -- a semi-major update with various enhancements and corrections. These include

- corrections to the main alignment display,
- improvements to the help system,
- major changes to the sequence print-out options,
-- including addition of a dotplot display (curtesy of DottyPlot),
-- a phylogeny tree display (curtesy of TreeDraw Deck & J. Felsenstein's DrawTree),
-- improved Pretty Print, which now has a single sequence form and a better aligned sequence form,
-- improved Restriction map display,
- addition and updating of several e-mail service links,
-- including Blast Search and Genbank Fetch via NCBI,

-- BLOCKS, Genmark, and Pythia services,
- updated Internet gopher client (equal to GopherApp),
- editable Child Tasks dialogs
- addition of links to Phylip applications as Child Tasks
- addition of Phylip interleaved format as sequence output option

11 June 92, version 1.6a35 is primarily a bug fix release. Several of the disasterous bugs have been squashed. This version now works on the Mac SE model, except for sendmail. No new features have been added.

7Jun92, v. 1.5a?? -- fixed several of the causes of mysterious bombs (mostly uninitialized handles), link b/n multiseq and 1-seq views is better now, folded in GopherApp updates, death date moved to Jan 93,

25Mar92, v1.5a32 (or later). First release to general public. Includes Internet Gopher client. Also released subset as GopherApp for non-biologists.

4Mar92, v 1.4a38 -- added base sliding in align view. Bases now slide something like beads on an abucus. Select a section with mouse, then grab section and shift left or right. Gaps are inserted/removed as needed. For use as contig aligner, still needs equivalent of GCG GelOverlap to automatically find contig/fragment overlaps.

Also added "Degap" menu item, to remove "." and "-". Fixed several small bugs including Align pretty print which again should display.

2Mar92, v 1.4a19 -- fixed several annoying bugs, see SeqApp.Help, section on bugs for their resolution. These include Complement/Reverse/Dna2Rna/ Translation which should work now in align view; Consensus menu item; entering sequence in align window now doesn't freeze after 30+ bases; pearson/fasta format reading; ...

10Feb92, v 1.4a6 -- fix for Mac System 6; add Internet service dialogs for Univ. Houston gene-server, Geneid @ BU, Grail @ ORNL; correct About Clustalv attribution.

5Feb92, v 1.4a4 -- limited release to network resource managers, clustalv authors, testers.

Vers 1.4, Dec91 - Feb92. Dropped multi-sequence picker window, made multi-align window the primary view (no need for both; extra confusion for users). added pretty print, restriction map, sequence conversions. Generalized "call clustal" to Hypercard-like, System 7 aware menu for calling external tasks. Fleshed out internet e-mail objects, added help objects, window menu, nucleic/amino help windows. Many major/minor revisions to all aspects to clean out bugs. Preliminary release to a limited set of testers (1.4a?)

Vers. 1.3, Sept - Dec91. Modified clustalv for use as external app (commandline file, background task, ...). Added basic Internet e-mail routines call clustal routine (preliminary child task) Many major/minor revisions to all aspects to clean out bugs.

Jun91-Aug91: overwork at other tasks kept SeqApp on back burner.

Mar91-Jun91: not much work on SeqApp, fleshed out TCP methods (UTCP, USMTP, UPOP).

Feb 1991, vers 1.2? made available to Indiana University biologists and NCBI biocomputists.

Vers. 1.1, Oct 1990, multiple sequence picker and multiple sequence alignement window, including colored bases, added to deal with alignment and common multi-sequence file formats.

Version 1, Sep 1990. Single sequence edit window + TextEdit window, from MacApp skeleton/example source + readseq.