Accelrys GCG

User Release Notes

VERSION 11.0 JUNE 2005

Copyright

(1) Copyright

(2) Trademark

(3) Restrictions on Government Use

This is a “commercial” product. Use, release, duplication, or disclosure by the United States Government agencies is subject to restrictions set forth in DFARS 252.227-7013 or FAR 52.227-19, as applicable, and any successor rules and regulations.

(4) Acknowledgments and References

To print photographs or files of computational results (figures and/or data) obtained using Accelrys software, acknowledge the source in an appropriate format. For example:

> “Computational results obtained using software programs from Accelrys Software Inc. Dynamics calculations performed with the Discover program using the CFF forcefield, ab initio calculations performed with the DMol^3 program, and graphical displays generated with the Cerius^2 molecular modeling system.”

To reference an Accelrys publication in another publication, Accelrys Software, Inc., is the author and the publisher. For example:

> Accelrys, Inc., Cerius^2 Modeling Environment, Release 4.8, San Diego: Accelrys Software Inc., 2005.

(5) Request for Permission to Reprint and Acknowledgment

Accelrys may grant permission to republish or reprint its copyrighted materials. Requests should be submitted to Accelrys Scientific Support, either through electronic mail to support@accelrys.com, or in writing to:

> Accelrys Scientific Support
> 10188 Telesis Court, Suite 100
> San Diego, CA 92121

Please include an acknowledgement “Reprinted with permission from Accelrys Software Inc., Document name, Month Year, Accelrys Software Inc., San Diego.” For example:

> Reprinted with permission from Accelrys Software Inc., Cerius^2 User Guide, June 2005, Accelrys Software Inc.: San Diego.

Contents

User Release Notes

Initializing GCG

This is a two step process:

Step 1.

```
% source <path>/startup
```

The former “gcgstartup” has been renamed “startup”. A convenient alias “gcgstartup” is present for backwards compatibility. Note that “startup” is now present at the root of the installation.

Startup sets up the $GCGROOT environment variable, validates that the installation directories that are present and correct, adds the bin directory to the PATH environment variable, defines the “gcg” and “gcgsupport” aliases and sets up the “.wp” preferences directory if it does not exist.

Step 2.

```
% gcg
```

“gcg” is an alias that sources the “etc/aliases” file. This will display the familiar GCG banner and creates a large number of aliases. This includes the convenient aliases like “to”, “up”, “home” etc, and “noglob” versions of all of the GCG programs. (This is required so that you can use the wildcard character ‘*’ when specifying database sequences.)

New Directory Structure

GCG 11.0 has a significantly changed directory structure to bring it more in line with typical UNIX applications. From the installation root, the following directories are present;

- **bin** – the location of all the binary executable files and scripts
- **doc** – the html on-line help files and GCG documentation is stored here
- **etc** – this directory holds all the configuration files for GCG
- **lib** – contains the shared object libraries required by GCG programs
- **sbin** – contains system administration executable files and scripts
- **share** – contains various shared files, mostly data files for the algorithms (e.g. genetic codes, restriction enzyme files etc)
- **var** – contains account management, licensing and logging files.

New programs

SeqMerge

Provides an easy-to-use graphical interface for managing sequencing assembly projects and takes the place of the command-line GelAssemble programs. SeqMerge uses an X-windows interface to select, view, manipulate, assemble and disassemble alignments of multiple sequence fragments. SeqMerge is capable of handling large contig assemblies.

Plus (+) programs

Many GCG programs have been rewritten using a new UNIX-friendly framework. These new programs are easily identified as they always have a "+" character appended to their name, for e.g. BLAST+. For backwards compatibility, we have also retained the original non-plus versions of these programs; hence you will see both BLAST and BLAST+.

Characteristics of a plus (+) program

Unlimited sequence length – for the most part, the plus (+) programs have no restrictions on sequence length. However, some programs are limited by restrictions in third party code, for instance, the FastA+ program can only accept input sequences up to 32,000bp in length, a restriction of the underlying algorithm. However, FastA+ can search database sequences of any length, unlike the original Wisconsin Package 10.3 FastA.

Enhanced file format support – the new plus (+) programs accept files in a number of different formats, in addition to the SSF, MSF, RSF, and GCG database formats accepted by the non-plus applications. Currently supported additional file formats are: GenBank, EMBL, PIR, Swiss-Prot, FastA, and BSML.

Use of parameter aliases **–** the new programs uses short aliases for each parameter. You must type either the alias or the full parameter. You cannot user partial parameter names as with the legacy applications. You can continue to use <program `name> -check` to list all of the parameters and aliases available for a program.

The following are plus (+) programs that have the enhanced capabilities listed above and will have the functionality of the corresponding non-plus counterpart.

Data Formatting Utilities

DataSet+

DataSet+ creates a GCG data library from any set of sequences.

Database Searching

FastA+

FastA+ does a Pearson and Lipman search for similarity between a query sequence and a group of sequences of the same type (nucleic acid or protein). For nucleotide searches, FastA+ may be more sensitive than BLAST+.

BLAST+

BLAST+ searches one or more nucleic acid or protein databases for sequences similar to one or more query sequences of any type. BLAST+ can produce gapped alignments for the matches it finds.

NetBLAST+

NetBLAST+ searches for sequences similar to a query sequence. The query and the database searched can be either peptide or nucleic acid in any combination. NetBLAST+ can search only databases maintained at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA.

FindPatterns+

FindPatterns+ identifies sequences that contain short patterns like GAATTC or YRYRYRYR. You can define the patterns ambiguously and allow mismatches. You can provide the patterns in a file or simply type them in from the terminal.

SSearch+

SSearch+ does a rigorous Smith-Waterman search for similarity between a query sequence and a group of sequences of the same type (nucleic acid or protein). This may be the most sensitive method available for similarity searches. Compared to BLAST+ and FastA+, it can be very slow.

Phlygentic Analysis

PAUP+

Primer Selection

Prime+

Prime+ selects oligonucleotide primers for a template DNA sequence. The primers may be useful for the polymerase chain reaction (PCR) or for DNA sequencing. You can allow Prime to choose primers from the whole template or limit the choices to a particular set of primers listed in a file.

Mapping

Map+

Map+ maps a DNA sequence and displays both strands of the mapped sequence with restriction enzyme cut points above the sequence and protein translations below. Map can also create a peptide map of an amino acid sequence.

Sequence Retrieval

Fetch+

Fetch+ copies GCG sequences or data files from the GCG database into your directory or displays them on your terminal screen.

NetFetch+

NetFetch+ retrieves sequences from NCBI listed in a NetBlast+ output file. You can also use it to retrieve sequences individually by sequence name or accession number. The output of NetFetch+ is an RSF file.

Pattern Recognition

MEME+

MEME+ finds conserved motifs in a group of unaligned sequences. MEME+ saves these motifs as a set of profiles. You can search a database of sequences with these profiles using the MotifSearch program.

Protein Analysis

TransMem+

TransMem+ scans for likely transmembrane helices in one or more input protein sequences.

HTHScan+

HTHScan+ scans protein sequences for the presence of helix-turn-helix motifs, indicative of sequence-specific DNA-binding structures often associated with gene regulation.

CoilScan+

CoilScan+ locates coiled-coil segments in protein sequences.

SPScan+

SPScan+ scans protein sequences for the presence of secretory signal peptides (SPs).

Sequence utilities

SeqManip+

SeqManip+ is a utility program that allows the user to perform manipulations of sequences, including translation, back translation of protein sequences, and sequence splitting. SeqMinip+ replaces several separate utility programs of earlier releases.

SeqStat+

SeqStat+ is a utility program that reads through any number of input sequences and provides some basic statistics about the files, including total length, number of sequences, and average length. Additionally it provides some extended information about the sequences depending on their type (protein or nucleotide), such as G+C% content.

Database utilities

FormatDB+

Combines any set of GCG sequences into a database that you can search with BLAST.

Importing / Exporting

SeqConv+

SeqConv+ is a utility program that provides batch conversions between different sequence formats. To the program easily converts between file formats and easily import data into Accelrys' bioinformatics applications. In addition, the converter allows the user to convert our internally used formats (e.g. BSML, RSF) into formats more commonly accepted by third-party tools. The supported file formats will include BSML, GenBank, FastA, EMBL, and RSF.

Multiple Sequence comparison

ClustalW+

Creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments and is able to plot a tree showing the clustering relationships used to create the alignment.

Unsupported Programs of this release

The programs listed below are available but not supported for version 11.0 of GCG.

- Spew
- ShiftOver
- SetKeys
- SeqED
- Replace
- OneCase
- PepData
- LineUp
- ListFile
- Lprint
- GelAssemble
- GelDisassemble
- GelEnter
- GelMerge
- GelStart
- GelView
- GetSeq
- Compress Text
- ChopUp
- Detab
- ExtractPeptide
- Red

The below mentioned programs are deprecated. The functionality of these programs has been incorporated into a single program “SeqConv+”.

- ToFASTA
- FromEMBL
- FromFASTA
- FromGenBank

Known Issues

GCG 11.0: On all platforms

1. Map+ - Usage of -`Enzymes` parameter:

   In Map program, -`enzymes=A*` selects all the enzymes that start with 'A'. In Map+, - `enzymes=`^A is the equivalent of A* (in Map) that selects all enzymes that start with 'A'.

2. Map+ Options, which are not implemented in GCG 11.0:

   i. Open (for mapping only the Open Reading Frames)

   ii. Lines per page

   iii. Display both forward and reverse cut positions[-`BOTtom`]

   iv. Display enzymes cut sites in vertical mode [-`VERtical`]

   v. Display cut sites by vertical line [-`NOCUTline`]

   vi. Display sequence in the output [-`NOSEQline`]

   vii. Display Numbered scale line in the output [-`NOSCALEline`]

   viii. Feature Character to separate each line in the map output [-`FeatureChar`]

3. Dataset+

   i. To avoid creating flat file databases from duplicate sequences, Wisconsin Package 10.3 database utilities supported an -exclude option, which allowed this user to specify a file containing accession numbers to exclude from loading. For data exclusion in GCG 11.0, the filename given as input to -exclude option should contain the <sequence name> to be excluded and not <accession number>.

   ii. If index is set to false, .offset and .seqcat files still get created. The expected behavior is that these files should not be created. However, these files need to be present, if you would like to create indexes later.

   iii. Dataset+ cannot handle sequence files having sequence name length greater than 20 characters.

   iv. Special characters like hyphen (-) and underscore(_) are valid characters and can be used as part of the logical name. Other special characters like #,@,!, etc are not supported. (All GCG programs fail to read from database library containing special characters (other than hyphen and underscore) in their library logical name.)

   For eg: globin#, globin* are not valid library logical names.

4. Blast+ program will not prompt for selection of databases. If you do not know your database choice, you must run Blast+ service twice. Run it first with the commandline:`% blast -dbreport`. And then the second time specifying the database to be searched. ..
5. The Name program cannot be used to set new shortcuts specifying directories in 11.0. However, it can be used for the display of all the possible shortcuts supported by GCG 11.0.
6. Symbol program is used to generate new symbols that are most frequently used by the users. In GCG 11.0, Symbols cannot be used to build new symbols, instead can be used to view the Symbols existing or already set in GCG 11.0 environment.
7. NetFetch+ output file does not display the accession numbers that were fetched at the beginning of the file as in Wisconsin Package 10.3 output.
8. SRS/Lookup indices cannot be created for Refseq entries
9. PrettyBox program fails to generate a readable post script format if the input file contains more than 76 sequences.
10. Batch jobs on all new plus (+) applications fail to send an email to the corresponding user(s).
11. New plus (+) programs do not display the option "add what to command line" with -`check` option. In other programs, this option can be used to specify additional parameters.
12. Interactive prompts do not show begin and end range of the input sequence(s) for all new plus (+) programs. The defaults are set to 1 and -1 for begin and end range respectively.
13. New plus (+) programs cannot generate plots, charts, or other graphics
14. Plotting parameters like width and height of the graphics cannot be configured for existing programs.
15. GCG 11.0 programs are unable to read input files that are greater than 2 GB.

SeqWeb 3.0, All Platforms

1. Plus (+) programs cannot generate plots, charts, or other graphics.
2. Plotting parameters like width and height of the graphics cannot be configured for plus (+) programs and existing programs.

GCG 11.0 on AIX

1. Motifs programs terminates abruptly with an error message:

```
exec(): 0509-036 cannot load program <Installed Path>bin/motifs
because of memory issues
```

2. Fasta_native programs have a limitation for large databases. The database search set should not exceed ~100,000 sequences.

GCG 11.0 on Linux

1. On RedHat 7.2, BLAST, BLAST+, and FormatDB+ do not work because NCBI BLAST version 2.2.9 is installed by default. You can extract blast-2.2.6_rh72.tgz (using the command *cd $GCGROOT/bin; tar-xvzf from CDROM/lp/blast-2.2.6_rh72.tgz*) library to downgrade these to 2.2.6 that work fine on RedHat 7.2.
2. Prior to building lookup indexes (gcg_srsbuild), please run *unlimit*, if you use csh.
3. On recent versions of RedHat Linux, if you get errors about *pthread_cancel*, please use your system's versions of standard libraries. You can do this by:

```
cd $GCGROOT/lib/
rm -f libgcc_s.so.1
rm -f libstdc++.so.5
ln -s /lib/libgcc_s.so.1 .
ln -s /usr/lib/libstdc++.so.5 .
```

GCG 11.0 on Irix

1. On Irix, the following programs do not work:

- peptidesort
- profilescan
- foldrna, mountains and domes

2. Following programs need more memory than other programs. For these to run, you need to set *unlimit* (if you use tcsh) or *ulimit* (if you use ksh/bash) before execution.

- motifs
- mfold
- hmmerpfam

SeqWeb 3.0 on AIX

1. Motifs program fails to run when the parameter 'Number of Mismatches' is set to '2'.
2. The following programs fail to run when invoked from a SeqWeb 3.0b installation on AIX platform.
   i. MotifSearch program
   ii. Profile Search program
3. After successful installation of GCG 11.0 and SeqWeb 3.0 software packages on AIX, a License check error is displayed when programs like Blast+ are invoked.

4. Run time of a Job in the Job Manager window is incorrectly displayed as 10:00:00 instead of 00:00:00 while the job is still running, but the final run time displayed after the job is completed is correct.
5. After successful installation of the SeqWeb 3.0b on AIX, the SeqWeb Administrator fails to login with the administrator User ID/Passwd. The system displays ‘Authorization required’ error message.

SeqWeb 3.0 on Irix

1. If you need to run the programs listed under *GCG 11.0* on Irix, as having higher memory requirement, you need to increase the memory limits to the user under which SeqWeb 3.0 is run. Please contact your Irix system administrator about information on how to do this.
2. The following programs fail to run when invoked from a SeqWeb 3.0b installation on Irix platform.
    - PeptideSort program
    - ProfileScan program
3. Run time of a Job in the Job Manager window is incorrectly displayed as 10:00:00 instead of 00:00:00 while the job is still running, but the final run time displayed after the job is completed, is correct.

Package-wide bug fixes

The following bugs have been fixed in GCG 11.0.

- In any GenBank record, the feature location was truncated if the feature has multiple lines and also the order of join and compliment are reversed while parsing the record. This has been fixed in the GCG framework in this release.
- GCG startup script is fixed to set the plot size variables correctly for GIF and PNG drivers
- GCG Framework handles output files generated by programs in a different manner than in Wisconsin Package 10.3. The output files are suffixed with _1 , _2 for every run of the input, so that they are not overwritten.
- Migration scripts such as migratedbs.sh, migratepersonaldbs.sh are modified to enable the migration necessary configuration files from Wisconsin Package 10.3 version to GCG 11.0 version.
- GCG programs fail when UniProt sequences has DOI Feature tag. This is fixed in the GCG framework.
- GCG FastA code has been fixed for parameters such as begin and end range specification and database qualifiers at the command line.

- Dataset program was unable to accept particular EMBL Sequences. This has been fixed in the GCG framework.
- SeqLab failed to display the Graphic features within the SeqLab Editor for sequences that were retrieved from SeqStore.
- WPD (GCG 11.0 Daemon) Security has been enhanced.
- Importing sequences into SeqLab was abruptly terminating for certain sequence records of GenBank. This has been fixed.
- SeqLab now throws an appropriate error message when we try to load any improper list files, lacking a list file header.
- GCG Programs now accept -`def` as an alias for –`default.`
- SeqLab can now select and upload multiple files properly, through the "multiple file selection window".
- MOTIFS no longer crashes on certain prosite entries.
- SeqMerge no longer fails to upload large SCF files through SeqLab.
- Stdin (Standard input) could not be used with some of the Wisconsin Package programs such as Dataset. Users can now use stdin like -in=- to input data via stdin.
- Comments tag (CC) of some EMBL-format files not read properly by GCG programs. This is fixed in the GCG framework.
- GCG programs failed to handle sequence files with spaces in the file name.
- GenbanktoGCG no longer fails for sequence records that have lines greater than 86 characters.