

ALLENMINER users guide v1.03



Fred P. Davis, HHMI-JFRC
davisf@janelia.hhmi.org
<http://research.janelia.org/davis/allenminer>

November 10, 2009

Abstract

This document describes how to install and use ALLENMINER to perform custom queries on three-dimensional Allen Brain Atlas data.

1 Introduction

ALLENMINER is a software package that allows expression queries of user-specified regions of the mouse brain, using the 3D XPR files available from the Allen Brain Atlas. This tool allows for the identification of genes that are (1) expressed, (2) expressed specifically, (3) expressed non-uniformly, or (4) expressed in a graded fashion, in a user-specified region of interest of the mouse brain. The region of interest can be specified as a set of ABA reference brain region names or as x,y,z bounds of a cuboid region of interest. (details below in Running ALLENMINER)

In addition, the code includes several routines that are generally useful for dealing with ABA XPR files. For example, the library includes code that can convert the ABA XPR format to a PDB (protein data bank) format file that can be viewed in any protein structure viewer, such as PyMOL. This is useful for visualizing ALLENMINER ROI definitions or for visualizing XPR files on platforms, such as GNU/Linux, where the ABA BrainExplorer is not available.

2 Downloading ALLENMINER

The package is freely available at <http://research.janelia.org/davis/allenminer>. The main software file is `allenminer_v1.03.tar.gz`. The `allenminer_manuscript_data.tar.gz` package contains files that are useful for ALLENMINER runs, including the `all_regions.roi_list.out.gz` file that accelerates ABA brain region searches.

3 Installation

3.1 Prerequisites

1. Perl, Python

ALLENMINER requires a functional installation of both Perl (<http://www.perl.org>) and Python(<http://www.python.org>). Both of these programs are usually installed by default on all Mac OS X and GNU/Linux machines. Windows users can install them using the Cygwin package (<http://www.cygwin.com>).

2. Bit::Vector perl module from CPAN: <http://search.cpan.org/search?query=bit-vector&mode=all>
3. XML::Twig perl module from CPAN - only necessary to create a local mirror of the ABA XPR files: <http://search.cpan.org/search?query=xml-twig&mode=all>

3.2 Local mirror of ABA XPR files

A local repository of ABA XPR (~8 GB) and XML (~250MB) files is only necessary for non-atlas ROI definitions or for gradient/patterning queries. It is not necessary to have the ABA XPR files for enrichment queries that refer to ABA regions. The `all_regions.roi_list.out.gz` file available from the `manuscript_data` package on the website quantifies the expression of all genes in each ABA atlas region. The "fast-query" run mode uses this file to perform ABA enrichment queries.

To setup a local repository, run:

```
perl src/setup_ABA_XML_XPR_mirror.pl
```

3.3 Installing ALLENMINER

1. Place the directory containing `AllenMiner.pm` in your `PERL5LIB` environment variable.
For example, if you run a `csh` or `tcsh` shell, add this to your `.cshrc` file:

```
setenv PERL5LIB {$PERL5LIB}:/your/allenminerpm/directory
```

For a `bash` shell, add this to your `.bashrc`:

```
PERL5LIB=/your/allenminerpm/directory:$PERL5LIB  
export PERL5LIB
```

2. Edit the `AllenMiner.pm` specs section (line 77-84) to point to your:
 - Allen Brain Atlas XPR file directory (see *Local mirror of ABA XPR files*)
 - ABA AtlasAnnotation100.sva file, available from the API at <http://mouse.brain-map.org/api/index.html>; This file is necessary to specify ROI using ABA brain regions.
 - ABA atlas brainstructures.csv file (also from the API package). This file is necessary to specify ROI using ABA brain regions.
3. Edit `SGE.pm` to reflect your local SGE computing cluster settings, if you want to run queries in parallel.

4 Running ALLENMINER

ALLENMINER has several run modes, eight of which we describe here.

4.1 run mode `roi_list`: Quantify the expression and enrichment of genes in an ROI

```
perl allenminer.pl -mode roi_list -roidef_fn ROI_DEFINITION_FILE [-cluster_fl 1] [-xpr_list_fn XPR_LIST_FILE]
```

- `-roidef_fn roifile`: file containing ROI definition (see *Input file formats* section)
- `-xpr_list_fn xpr_list_file`: file containing a list of XPR files to process;
if not specified, will parse all XPR files located in the `$specs->{allen_xpr_dir}` specified in `AllenMiner.pm`
- `-cluster_fl <0|1>` - optional
if 1 will run the query in parallel using an SGE computing cluster, as specified in `SGE.pm`
- `-xyz_details 1` - optional
if 1 will create text format tab-delimited files for each XPR file that contains detailed expression data for ROI coordinates.
- `-xyz_details_compress 1` - optional (for use with `-xyz_details 1`)
will compress `xyz_details` output files
- `-xyz_details_out_suffix SUFFIX` - optional
will append `xyz_details` output files with the suffix "SUFFIX"
- `-xyz_details_out_dir OUTDIR` - optional
will deposit `xyz_details` output files into the `OUTDIR` directory

4.2 run mode `fastquery_aba_ref_atlas`: Perform an index-accelerated query of ABA brain regions

```
perl allenminer.pl -mode fastquery_aba_ref_atlas -query_specs_fn QUERY_SPECS_FILENAME -aba_results_fn all_regions.roi_list.out.gz [-out_fn OUTPUT_FILE_NAME] [-outfile_prefix OUTPUT_FILE_PREFIX]
```

- `-query_specs_fn <QUERY SPECIFICATIONS FILE>`
this file describes the details of the query to be performed (see *Input file formats* section).
- `-aba_results_fn <ABA_INDEXING_FILE>`
specify the location of the `all_regions.roi_list.out.gz` file available in the `manuscript_data` package on the ALLENMINER website. This file contains the pre-computed expression. This file quantifies the expression of all genes in each ABA atlas region.

- `-outfile_prefix <OUTPUT FILE PREFIX>` (optional) specify this if you would like the results of each query to be displayed in a separate file, with this prefix. The suffix is “_QUERYNAME.allenminer.out”
- `-out.fn <OUTPUT FILE NAME>` (optional) specify this if you want the results of all queries to be displayed in the same output file

If nothing is specified about an output file, it displays the results in:
“aba_fast_query.QUERYNAME.PID.allenminer.out” where PID is the process id.

4.3 run mode `roi_partition`: **Generate new ROIs by partitioning an ROI along RC, ML, or DV axes**

```
perl allenminer.pl -mode roi_partition -roidef_fn
ROI_DEFINITION_FILE -roi_basename PARTITIONED_ROI_BASENAME_ -
numbins NUMBER_OF_BINS -axes <AP|ML|DV> [-fitted_cuts 1]
```

- `-roidef_fn ROIFILE`: file containing ROI definition (see *Input file formats* section)
- `-roi_basename BASENAME`
specifies the prefix of the name to give to the newly created ROI partitions.
- `-numbins NUMBER_OF_BINS`
specifies the number of ROI partitions to create
- `-axes <AP|ML|DV>`
specifies the axis along which to partition the ROI; AP = anteroposterior (or RC = rostral-caudal), ML = mediolateral, DV = dorsoventral.
- `-fitted_cuts 1`
optional flag that specifies that ROI partitions should be generated with the space across the partition axis equally split among the partitions. This results in partition edges that adapt to the shape of the ROI. If this option is not specified, the partitions edges are parallel to the axis.

4.4 run mode `calcentropy_xpr_roi_results` - **compute the gradient and patterning scores from the results of a `roi_list` run**

```
perl allenminer.pl -mode calcentropy_xpr_roi_results -results_fn
ROI_LIST_OUTPUT_FILE -out_fn OUTPUT_FILE_NAME
```

- `-results_fn ROI.LIST.RESULTS.FILE`
specifies a file containing the output of an `roi_list` ALLENMINER run.
- `-out.fn OUTPUT_FILE_NAME`
specifies a file to hold the output.

4.5 run mode roi2pdb: Convert an ROI definition file to PDB format

```
perl allenminer.pl -mode roi2pdb -roidef_fn ROI_DEFINITION_FILE -  
pdb_fn OUTPUT_PDB_FILENAME
```

- -roidef_fn ROIFILE: file containing ROI definition (see *Input file formats* section)
- -pdb_fn OUTPUT_PDB_FILENAME: name of PDB output file

4.6 run mode xpr2pdb: Convert an XPR file to PDB format

```
perl allenminer.pl -mode xpr2pdb -xpr_fn XPR_filename -pdb_fn  
OUTPUT_PDB_FILENAME
```

- -xpr_fn XPR_FILE_NAME - name of ABA XPR file to convert
- -pdb_fn OUTPUT_PDB_FILENAME: name of PDB output file

4.7 run mode xpr2txt: Convert an XPR file to text format

```
perl allenminer.pl -mode xpr2txt -xpr_fn XPR_filename -out_fn  
OUTPUT_FILENAME
```

- -xpr_fn XPR_FILE_NAME - name of ABA XPR file to convert
- -out_fn OUTPUT_FILENAME: name of output file

4.8 run mode estimate_roi_sampling: Estimate the number of coronal and sagittal slices sampled across an ROI

```
perl allenminer.pl -mode estimate_roi_sampling -roidef_fn  
ROI_DEFINITION_FILE
```

- -roidef_fn ROIFILE: file containing ROI definition (see *Input file formats* section)

4.9 Runs described in the manuscript

Several enrichment, patterning, and gradient queries are described in the manuscript describing ALLENMINER (See *Citing ALLENMINER* section). The associated input and output files are available in the allenminer_manuscript_data.tar.gz file on the website.

- Collate the expression in the ROI defined in neocortex.roi in the XPR files specified in xpr_file_list.txt, using the compute cluster.

```
perl allenminer.pl -mode roi_list -roidef_fn neocortex.roi -
cluster_fl 1 > neocortex.roi_list.out
```

- Calculate the entropy across the ROI bins in the roi_list result file neocortex_5APbins.roi_list.out and send the results to the file neocortex_5APbins.entropy

```
perl allenminer.pl -mode calentropy_xpr_roi_results -results_fn
neocortex_5APbins.roi_list.out -out_fn neocortex_5APbins.
entropy
```

- Convert ROI defined in neocortex.roi into PDB format neocortex.roi.pdb

```
perl allenminer.pl -mode roi2pdb -roidef_fn neocortex.roi -
pdb_fn neocortex.roi.pdb
```

- Create 5 fitted mediolateral bins from the ROI defined in cp.roi; name the new ROIs using the base CP_5MLbins_

```
perl allenminer.pl -mode roi_partition -roidef_fn cp.roi -
roi_basename CP_5MLbins_ -numbins 5 -axes ML -fitted_cuts 1
```

5 Input file formats

5.1 ROI definition file

This file specifies one or multiple ROI. Examples are provided in `example/*.roi`

It is a tab-delimited file with the following fields:

1. ROI name - required
2. ROI subname - optional, leave blank if don't want to use it
3. type of definition - one of: point, brainstrx, boxbounds
4. if point: x,y,z in Allen BrainExplorer (and .XPR) coordinates
if brainstrx: either a single abbreviation from Allen Brain Atlas (eg CTX) or if want to specify a subset, as follows: CTX=on,OLF=off,HPF=off
if boxbounds: one of xmin, xmax, ymin, ymax, zmin, or zmax - if an axes bound is not defined, extends to edge of axes.
5. (if boxbounds): number value or equation defining the axis boundary
 - eg, '10' to specify a hard upper axis limit of 10
 - eg, 'y * 0.5' to specify a y-dependent axis limit of half the y value

If multiple lines are specified with the same ROI name, they get appended onto the ROI definition.

5.2 Query specs file for fastquery_aba_ref_atlas run mode

The query_specs file is a tab-delimited file with the following fields:

1. query name: arbitrary label for labeling the output. Ideally don't include spaces.
2. query type: either 'enrichment' or 'level' depending on whether you would like a ROI enrichment score or expression level
3. roi1: ABA atlas region name (eg, CP for caudoputamen)
4. roi2: specify a second ROI - only required if this is an enrichment query: will compute the relative enrichment of each gene in roi1 vs roi2.

5.3 xpr_list_fn

This is simply a 1 column list of XPR files to process.

6 Citing ALLENMINER

A tool for identification of genes expressed in patterns of interest using the Allen Brain Atlas. Fred P. Davis and Sean R. Eddy. manuscript submitted

7 Contact Information

Fred P. Davis
Howard Hughes Medical Institute
Janelia Farm Research Campus
19700 Helix Dr
Ashburn, VA 20147, USA
email: davisf@janelia.hhmi.org
phone: (571)-209-4000 x3037