Analyzing large data with Mappable Vector Library

Vladimir Dergachev

Introduction

Mappable Vector Library (MVL) is a file format optimized for memory mapping. It is designed for ease of use by both low-level C code and high-level scripting languages (such as R).

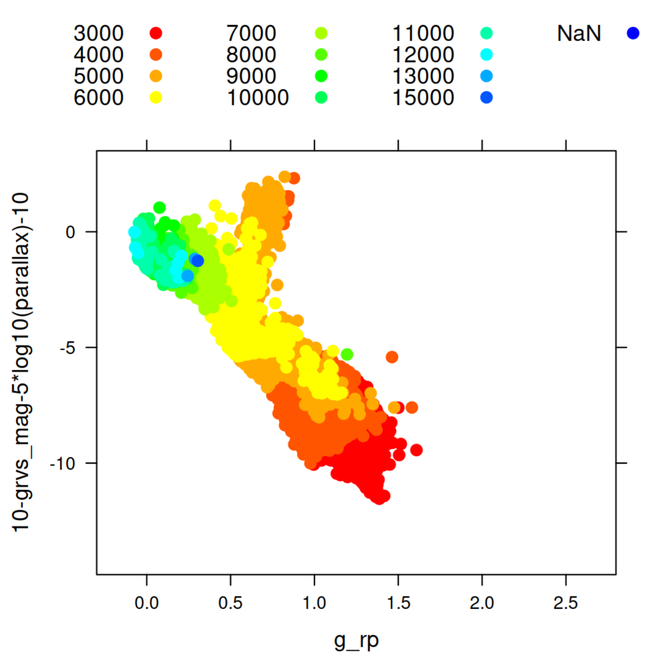
You can use MVL files to:

- Analyze as much data as fits on your solid state drive
- Exchange bulk binary-level data between C and R programs
- Share large data between multiple processes running on the same computer

There are two libraries for C (libMVL [1]) and R (RMVL [2]) providing seamless array-style access as well as database functionality of sorting the data and creating hash-based value and spatial indices.

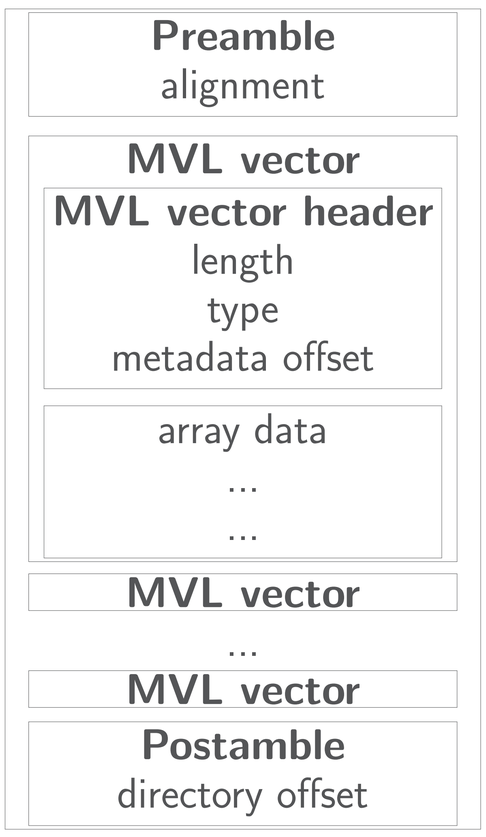
The library functions have been optimized to reduce the number of writes to solid state drives.

A galaxy in a file



The plot above is a variant of H-R diagram produced with Gaia data. The color of points reflects grouping using `teff_gspphot` variable. The X axis is a proxy for the color using `g_rp`, while the Y axis expression is an estimate of absolute magnitude. The code used to produce the plot is available at [3].

MVL file layout



Gaia data in MVL format

The main Gaia data is over 1 TB in size and is provided in 3386 compressed CSV files. While one can use online databases to search for a given object, it would be nice to explore Gaia data as a whole. To make this possible Gaia data was converted to MVL format [4] (available for download using HTTP and Torrent).

The main Gaia data in MVL format is 1.3 TB which is easily memory mapped into R. If your solid state drive is capable of 3.5 GB/s read speed (as many do), you can scan the entire data in under 7 minutes. Though, in practice, you probably will not need all the different variables.

For more sophisticated access there are indices allowing to search and scan in order of distance from Earth, or by sky position.

Gaia example

```
library("RMVL")
M<-mvl_open("gaia_dr3.mvl")
Mpi<-mvl_open("parallax_index.mvl")
# Find closest one million stars
idx<-Mpi$parallax_index[1:1000000]
plot(M$gaia[idx, "parallax"], log="xy")
```

MVL architecture

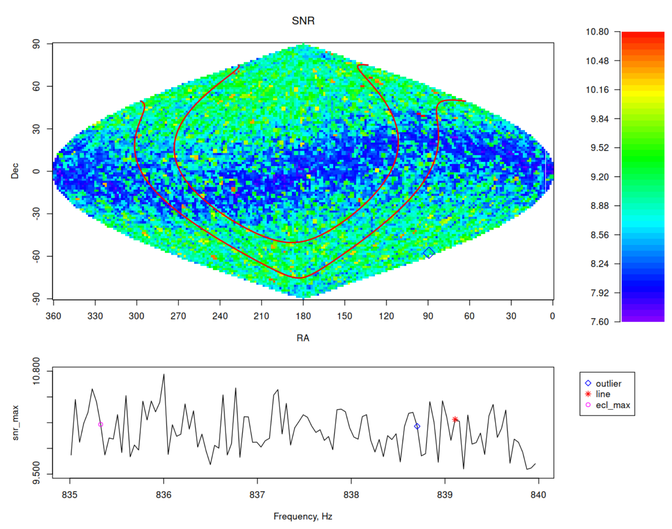
Unlike application-level databases the MVL files are designed for direct access to data. The basic element of MVL file is an MVL vector composed of a header, describing type and length of the data, followed by a linear array of data elements. The array is aligned to a configurable boundary, 64 bytes by default. This allows to use the data as is with vector arithmetic operations.

Each MVL vector is uniquely identified by its offset from the beginning of the MVL file. This makes it very easy to access the data in C - just memory map the file and use the offset to obtain a pointer to data. The offsets and array lengths are 64-bit.

The MVL vectors support character, integer and floating point types. In addition, there is a special *offset* vector that stores a list of offsets to other vectors. This allows to organize the data into a tree with arbitrary complexity and maps nicely to the internal representation used by scripting languages such as R or Tcl. Each MVL vector header can have an optional offset to metadata. A particularly useful metadata field is a vector of character names associated to elements of MVL vector array. This allows to create named lists, which can be used to represent sets, dictionaries and structures.

The MVL file ends with a postamble containing an offset to the top-level vector of offsets called a *directory*.

Gravitational wave atlas



All-sky searches for continuous gravitational waves sources produce multi-dimensional data sets. The plot above shows a slice of recently released atlas of gravitational wave sky [5] which data is parameterized by sky location and frequency band of potential sources in MVL format. Queries by sky position can use a spatial index.

References

[1] https://github.com/volodya31415/libMVL

[2] https://cran.r-project.org/package=RMVL

[3] https://github.com/volodya31415/gaia_mvl

[4] Gaia data in MVL format https://www.atlas.aei.uni-hannover.de/work/volodya/Gaia_dr3/

[5] Continuous gravitational wave atlas in MVL format https://www.atlas.aei.uni-hannover.de/work/volodya/O3a_atlas/

CONTINUOUS GRAVITATIONAL WAVES