

---

Subject: Re: Locating sequence of bytes within binary file  
Posted by [Craig Markwardt](#) on Wed, 16 Jun 2010 13:46:07 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

On Jun 15, 7:30 am, medd <med...@googlemail.com> wrote:

> Hi,  
>  
> I need to locate a given sequence of bytes within a binary file. I do  
> not manage to do it efficiently, and I wanted to ask if somebody here  
> has a clue.  
>  
> I saw that there are no functions in IDL to look for a given sequence  
> within a byte array, but there are very powerful functions to look for  
> a sequence within a string using regular expressions. This is what I  
> tried:  
>  
> fcontent = BYTARR((FILE\_INFO(fn)).size, /NOZERO) ;Variable where to  
> read in the file  
> OPENU, unit, fn, /GET\_LUN;, /SWAP\_ENDIAN  
> READU, unit, fcontent  
> IF(STREGEX(String(fcontent), String(sequence\_searched)) LT 0) THEN  
> print, 'sequence not found'  
>  
> This works!! ... But only as long as the file does not contain a byte  
> with the value 0 (which, too bad!, it does...)  
>  
> After looking a while, I found in this forum (message "Null terminated  
> strings") and in the IDL help that a string is truncated as soon as  
> this value is found. This explains why this method fails. But it does  
> not propose solutions... :(  
>  
> Do you know some smart workaround? Or do you know other efficient ways  
> in IDL to locate a sequence of bytes within a binary file?

You can use FFT cross-correlation to search for matching segments.

```
;; Sample byte data
haystack = byte(randomu(seed,1000000)*255)

;; This is the search string to be found
needle = haystack(12345:12444)

;; Cross-correlation from the IDL astronomy library
cc = convolve(haystack+0.,needle+0., /correl)
```

Then look for correlation peaks. At that stage, once you have identified candidate peaks, you can do a refined search to make sure you have an exact match. The peak will be located at the center of

the string, not the beginning.

I hadn't thought of this before, but this gives a way to do fuzzy matching because the correlation technique does not require exact numerical match at every point. However, this mostly works for longer search strings.

Good luck,  
Craig

---