

Re: Regular Expression – Matching Multiples of 3 Characters exactly.

# Re: Regular Expression – Matching Multiples of 3 Characters exactly.

---

*Source:* <http://coding.derkeiler.com/Archive/Python/comp.lang.python/2008-04/msg03829.html>

---

- *From:* blaine <[frikker@xxxxxxxx](mailto:frikker@xxxxxxxx)>
  - *Date:* Sun, 27 Apr 2008 19:31:42 -0700 (PDT)
- 

On Apr 27, 10:24 pm, castiro...@xxxxxxxx wrote:

On Apr 27, 8:31 pm, blaine <[frik...@xxxxxxxx](mailto:frik...@xxxxxxxx)> wrote:

Hey everyone,  
For the regular expression gurus...

I'm trying to write a string matching algorithm for genomic sequences. I'm pulling out Genes from a large genomic pattern, with certain start and stop codons on either side. This is simple enough... for example:

```
start = AUG stop=AGG  
BBBBBBAUGWWWWWAGGBBBBBB
```

So I obviously want to pull out AUGWWWWWAGG (and all other matches).  
This works great with my current regular expression.

The problem, however, is that codons come in sets of 3 bases. So there are actually three different 'frames' I could be using. For example:  
ABCDEFGHIJ  
I could have ABC DEF GHI or BCD EFG HIJ or CDE FGH IJx.... etc.

Re: Regular Expression – Matching Multiples of 3 Characters exactly.

So finally, my question. How can I represent this in a regular expression? :) This is what I'd like to do:  
(Find all groups of any three characters) (Find a start codon) (find any other codons) (Find an end codon)

Is this possible? It seems that I'd want to do something like this: `(\w\w\w)+(AUG)(\s)(AGG)(\s)*` – where `\w\w\w` matches EXACTLY all sets of three non-whitespace characters, followed by AUG \s AGG, and then anything else. I hope I am making sense. Obviously, however, this will make sure that ANY set of three characters exist before a start codon. Is there a way to match exactly, to say something like 'Find all sets of three, then AUG and AGG, etc.'. This way, I could scan for genes, remove the first letter, scan for m