# Regular Expressions

Regular expressions are a very powerful part of your toolkit. They are meant to make searching text, finding patterns, extracting patterns automatic. I almost said "easy", but I caught myself because they can be extremely confusing and frustrating when they get complicated. Often called "regexes", regexes often make grown programmers cry.

## FASTA Files

Mucho biological information is encoded as simple text files. Sequences are often encoded as FASTA files:

*Examples of FASTA file needed here*

- Short reads from a sequencer may come as a FASTA file (more often as a FASTQ file, technically, but they're very similar). *Example* Here we have lots and lots of short sequenes. The "tags" start with `>` and have some weird identifying code the sequence is all the letters on the next lines, until you hit another `>` or get to the end of the file.
- Entire genomes often come as FASTA files. *Example, maybe human genome* In this case, the tags are named by chromosome, and the sequences are very long as each sequence represent an entire chromosome.

## Biology for a moment

- Transcription (making RNA from the DNA template) happens at specific positions along a chromosome
- RNA polymerase (a protein that does the transcription) binds to the beginning of a gene. But how does RNA polymerase know where to bind along the DNA?
- Super complicated topic, especially in eurakyotes (remember: organisms with nuclei)
- Promoter regions are special DNA sequences upstream of genes that RNA polymerases (yes there are different varieties) bind to, to start transcription. (It get more complicated than that, what with enhancers and repressors (??) but that's the quick verison)
- About a quarter of human genes contain a DNA sequence called a "TATA-box", with the DNA sequence `TATAAA`

## Find TATAAA in human genome

- Bring up human genome FASTA file. Someone please point out the first few lines with

TATA boxes

- Unreasonable, right? That's why we have computers as slaves
- Do magic:

```
grep 'TATAAA' <human genome>.fa
```

- Or some variant that shows TATA box in relativily short frame
- Maybe there's some variant of `TATAAA` where we can use alternate matching??

## Anchors: Count number of sequences in a FASTA file

- Introduce start anchor: `grep -c '^>' file.fa`
- What's a FASTQ file?
- Count number of sequences in a FASTQ file
    - Introduce end anchor: `$`
    - `grep -c '^+$' file.fq`

## | option, search for one of two promoters?

## Some example of range

## Some example of shortcut like \s \d etc

## The . wildcard

## Repeats: *, +, and whatever the other one is (not that * is NOT a wildcard for regexes)

## regex101 and/or pythex

## Regexes in Python. Oh boy.

This is where we introduce capturing