

# Output Nucleotide Subsequences with Context

Consider nucleotide sequences which contain the possible base values: 'A', 'C', 'G', 'T', and the end-of-sequence value: 'ε'.

**Implement a command-line program that accepts a nucleotide sequence on `stdin`, finds target nucleotide subsequence, and writes each subsequence to `stdout` along with its surrounding context.**

More specifically: given a nucleotide sequence  $S$ , a target subsequence  $T$ , and two numbers  $x$  and  $y$ ; for each  $T$  in  $S$ : print the  $x$  preceding nucleotides, the target  $T$ , and the  $y$  succeeding nucleotides.

An example command line

```
echo "ACACGTCAε" | matchseq -T:ACGT -x:1 -y:2
```

would yield

```
C ACGT CA
```

where C is the  $x=1$  preceding nucleotide, ACGT is the target nucleotide subsequence  $T$ , and CA is the  $y=2$  succeeding nucleotides.

Be aware that:

- $x$  may be zero, which indicates that no preceding nucleotides should be printed. Likewise,  $y$  may be zero, which indicates that no succeeding nucleotides should be printed. However,  $T$  will not be empty.
- The end-of-sequence value 'ε' will not appear in  $T$ .
- If the sequence contains fewer than  $x$  nucleotides before  $T$ , or fewer than  $y$  nucleotides after  $T$ , print as many as there actually are.
- Targets may overlap in the sequence  $S$ , and each should be treated as a distinct occurrence with its own preceding/succeeding context.
- `stdin` is a potentially unlimited stream, so be sure to consider the case where the size of  $S$  exceeds a system's memory.

## Example

```
echo "AAGTACGTGCAGTGAGTAGACCTGACGTAGACCGATATAAGTAGCTAε" | \
matchseq -T:AGTA -x:5 -y:7
```

should print the following lines (whose targets we've bolded):

```
A AGTA CGTGCAG
CAGTG AGTA GTAGACC
TGAGT AGTA GACCTGA
ATATA AGTA GCTA
```

Notice that the 2nd and 3rd lines display overlapping targets and that the 1st and 4th show fewer than  $x$  and  $y$  elements of context, respectively.