```
#!/bin/bash
# File: fasta2mer_sorc.sh   (Convert FASTA sequence into N-mers: fixed
at N=5)
# Usage: 'fasta2mer_sorc.sh in.fasta'
# Copyright 2019 Sage-N Research, Inc. All rights reserved. Licensed
for SORCERER.

tmp_seqs=$(mktemp /tmp/fasta2mer_seqs.XXXXX)  # Bare protein sequences
one per line
tmp_fwd1=$(mktemp /tmp/fasta2mer_fwd1.XXXXX)  # Forward-sequence 5-
mers starting at position 1..5
tmp_fwd2=$(mktemp /tmp/fasta2mer_fwd2.XXXXX)
tmp_fwd3=$(mktemp /tmp/fasta2mer_fwd3.XXXXX)
tmp_fwd4=$(mktemp /tmp/fasta2mer_fwd4.XXXXX)
tmp_fwd5=$(mktemp /tmp/fasta2mer_fwd5.XXXXX)
tmp_fwd6=$(mktemp /tmp/fasta2mer_fwd6.XXXXX)
tmp_fwd01to06=$(mktemp /tmp/fasta2mer_1to6.XXXXX)
tmp_fwd07to12=$(mktemp /tmp/fasta2mer_7to12.XXXXX)
tmp_fwd01to12=$(mktemp /tmp/fasta2mer_1to12.XXXXX)

# Within input from stdin, strip headers and concatenate subsequences
awk '{printf $1~/^>/?"\n":$0} END{printf "\n"}' $1 > $tmp_seqs  # NL
iff header, else print subsequence; note blank first line

# Get 12-mers with successive starting positions
cat                   $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd1
sed 's/.//'           $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd2
sed 's/..//'          $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd3
sed 's/...//'         $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd4
sed 's/....//'        $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd5
sed 's/.....//'       $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd6
cat $tmp_fwd1 $tmp_fwd2 $tmp_fwd3 $tmp_fwd4 $tmp_fwd5 $tmp_fwd6 |sort
|uniq > $tmp_fwd01to06
#
sed 's/......//'      $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd1
sed 's/.......//'     $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd2
sed 's/........//'    $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd3
sed 's/.........//'   $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd4
sed 's/..........//'  $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd5
sed 's/...........//' $tmp_seqs |fold -w12 |sort |uniq > $tmp_fwd6
cat $tmp_fwd1 $tmp_fwd2 $tmp_fwd3 $tmp_fwd4 $tmp_fwd5 $tmp_fwd6 |sort
|uniq > $tmp_fwd07to12
# Omit I (L isotomer) and unused chars; keep only 12-mers
cat $tmp_fwd01to06 $tmp_fwd07to12 |egrep -v 'B|I|J|O|U|X|Z' | awk
'length($1)==12' > $tmp_fwd01to12
rev $tmp_fwd01to12 | cat - $tmp_fwd01to12 |sort |uniq > s12mer.lst
#
rm $tmp_seqs $tmp_fwd1 $tmp_fwd2 $tmp_fwd3 $tmp_fwd4 $tmp_fwd5
$tmp_fwd6 $tmp_fwd01to06 $tmp_fwd07to12 $tmp_fwd01to12
```

```
# Successively one shorter
tmp1=$(mktemp /tmp/fasta2mer_1.XXXXX)
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s12mer.lst > $tmp1; sort $tmp1 |uniq > s11mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s11mer.lst > $tmp1; sort $tmp1 |uniq > s10mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s10mer.lst > $tmp1; sort $tmp1 |uniq > s09mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s09mer.lst > $tmp1; sort $tmp1 |uniq > s08mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s08mer.lst > $tmp1; sort $tmp1 |uniq > s07mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s07mer.lst > $tmp1; sort $tmp1 |uniq > s06mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s06mer.lst > $tmp1; sort $tmp1 |uniq > s05mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s05mer.lst > $tmp1; sort $tmp1 |uniq > s04mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s04mer.lst > $tmp1; sort $tmp1 |uniq > s03mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s03mer.lst > $tmp1; sort $tmp1 |uniq > s02mer.lst
awk '{printf("%s\n%s\n", substr($1,1,-1+length($1)), substr($1,2))}'
s02mer.lst > $tmp1; sort $tmp1 |uniq > s01mer.lst
rm $tmp1
```