# NAME

```
motifCounts.pl
```

# SYNOPSIS

% perl motifCounts.pl w x y z

w = <minimum LMER size>

x = <maximum LMER size>

y = < all | genic | intergenic >

z = < chr = chromosomes | plasmids >

e.g. % perl motifCounts.pl 4 4 all chr

means 4mers, whole genome, and chromosomes

# DESCRIPTION

## SUMMARY

This script assumes that the script cutter.pl has already been run. This script goes through all the files created by cutter.pl that match the type of data specified in the command line, counts the number of times each unique lmer appears in the genome as well as its reversed complementary sequence, and outputs the results into a series of .xls files one for each combination of lmer size and input file

## Input

### Genomic .fna files

This script takes input in the form of .fna files created by the script cutter.pl and stored up one level in the directory ../1_cutter/split_texts. The fna files are text files only(primarly?) containing the letters A T C and G

### Command line Arguments

The script takes two numbers to get the range of LMER sizes to use to count motifs, as well as what portion of the genome to read, which may be entire genome, the genic portions, or the intergenic portions and finally whether to read chromosomes or plasmids

### Output

The script stores the results as .xls files in the folder Cache_Counts, first sorting the data into a set of sub-directories by organism, and then another set of directories by LMER size within the sub directories for the specific organisms. The .xls file name is the name of the file read from then an underscore then the LMER size.xls

Ex: The results for an organism named Alcanivorax_borkumensis_SK2 using 4mers from the file NC_008260_chr01.fna would be stored in cache_counts\Alcanivorax_borkumensis_SK2\4mer in the file NC_008260_chr01_4.xls

The first line of the .xls file stores the percentage of G's and C's in the genome. Each successive line contains a specific motif, a tab and then the number of times that motif occurs.

Partial Example

```
 0.547277801049503
 AAAA 33490
 AAAC 26124
 AAAG 26812
 AAAT 22822
 AACA 25246
```

## AUTHORS

```
Mark D. LeBlanc
Donald W. Bass
```

## MODIFICATION HISTORY

**06/01/2010 (dwb)**

-- Updated code to be platform independent

**05/31/2010 (dwb)**

-- Added comments to code, and started pod documentation

**04/23/2009 (mdl)**

-- %GC added to top (header) of each output file of counts for each chunk thus, step 3_prepare4R will have to read this in first

**04/17/2009 (mdl)**

-- add %GC of each chunk into the vector(output file) of motif counts; calculated on the direct strand, based of course only on length of direct strand

**01/12/2009 (mdl)**

-- don't print the total number of motifs per file in first line; start with AAAA <count>, etc; this will make it consistent with the way we count in the NORMALIZED version

**01/07/2009 (mdl)**

-- make directory structure consistent with how we handle the normalized counts as calculated and stored by the lexomics:~mleblanc/DNA/2_countNnormalizeMotifs/allmers.sh (bash) script; essentially, store vector of counts for each chunk in a directory for each bug, e.g., /split_text/bug1__chr01/<chunk vectors>, /bug2__chr01/<chunk vectors> ...

**11/18/2008 (mdl)**

-- working with output from cutter; counting motifs in chunks

## COPYRIGHT INFORMATION