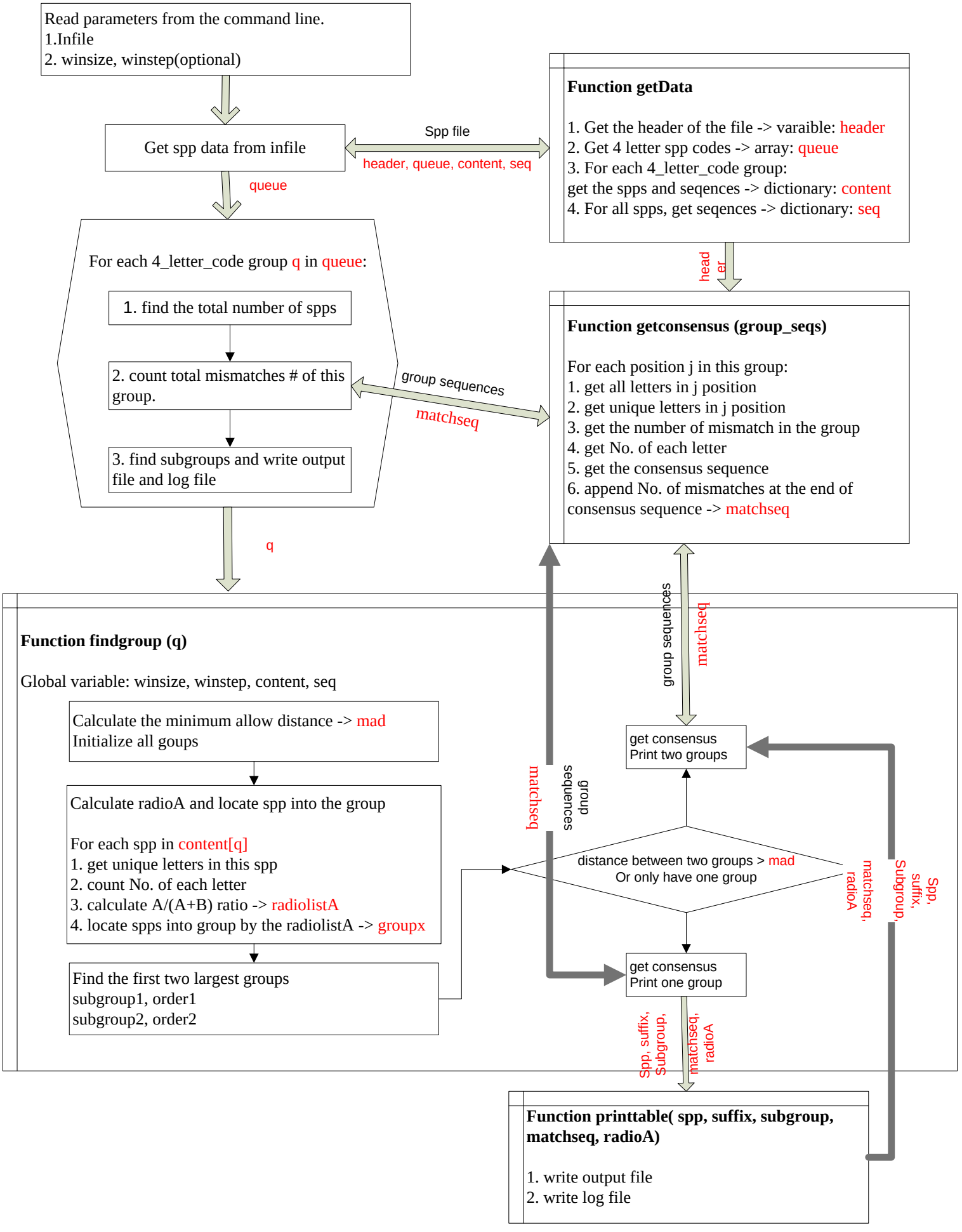
Read parameters from the command line.
1.Infile
2. winsize, winstep(optional)
Get spp data from infile
Spp file
header, queue, content, seq
queue
Function getData
1. Get the header of the file -> varaible: header
2. Get 4 letter spp codes -> array: queue
3. For each 4_letter_code group:
get the spps and seqences -> dictionary: content
4. For all spps, get seqences -> dictionary: seq
head er
For each 4_letter_code group q in queue:
1. find the total number of spps
2. count total mismatches # of this group.
3. find subgroups and write output file and log file
group sequences
matchseq
Function getconsensus (group_seqs)
For each position j in this group:
1. get all letters in j position
2. get unique letters in j position
3. get the number of mismatch in the group
4. get No. of each letter
5. get the consensus sequence
6. append No. of mismatches at the end of consensus sequence -> matchseq
q
group sequences
matchseq
Function findgroup (q)
Global variable: winsize, winstep, content, seq
Calculate the minimum allow distance -> mad
Initialize all goups
Calculate radioA and locate spp into the group
For each spp in content[q]
1. get unique letters in this spp
2. count No. of each letter
3. calculate A/(A+B) ratio -> radiolistA
4. locate spps into group by the radiolistA -> groupx
Find the first two largest groups
subgroup1, order1
subgroup2, order2
get consensus
Print two groups
matchseq
group sequences
distance between two groups > mad
Or only have one group
matchseq, radioA
Spp, suffix, Subgroup,
get consensus
Print one group
Spp, suffix, Subgroup,
matchseq, radioA
Function printtable( spp, suffix, subgroup, matchseq, radioA)
1. write output file
2. write log file