GENERAL INSTALLATION INSTRUCTIONS FOR CHAP

- Our software needs to run the RepeatMasker program, which can be obtained from http://www.repeatmasker.org/. You can modify the right-hand side of the line
  REPEATMASKER = RepeatMasker
  very near the start of the file conversion.sh to indicate the location of the RepeatMasker executable on your computer.
- If you want to use our Gmaj program to view the results, you will need to have Java installed on your computer.
- In the directory containing the unpacked versions of our programs, which we will call the "package directory", type "make" to compile our programs and install them in the bin subdirectory.
- For each gene cluster that you want to analyze, do the following.
  1. In the package directory, create a subdirectory for the cluster, which we will call the "cluster directory".
  2. In the cluster directory, create a subdirectory called "seq.d" and put your FastA-formatted sequence files in it, giving each file the appropriate species name, e.g., "human", "vervet".
  3. In the cluster directory, create another subdirectory called "exons.d" and put your gene annotation files (e.g., downloaded from cluster.bx.psu.edu) in it. These files should be named e.g. "human.exons", "vervet.exons", etc. An example of the format can be found at:
     http://globin.bx.psu.edu/dist/gmaj/gmaj_input.html#exon
  4. Put a Newick-formatted species tree (e.g. from cluster.bx.psu.edu) in the cluster directory. This file can have any name; for concreteness, let's suppose it is called "tree.txt".
  5. In the cluster directory (which contains subdirectories seq.d and exons.d, as well as the species tree, from steps 2-4), run the command:
     ../conversion.sh tree.txt
     The pipeline may run for more than an hour.
  6. Get a summary of the results from conversion.sh by running the command
     ../bin/gc-info all.gc
     and/or examine them in detail by running Gmaj with commands like
     ../gmaj.sh human
     or
     ../gmaj.sh vervet
  7. The file gmaj_geneconv.html provides a short tour of how to use Gmaj to investigate gene conversions, while more general documentation for Gmaj is available at:
     http:www.bx.psu.edu/miller_lab

DETAILS OF THE OUTPUT FORMAT

A summary of the results is printed at the end of the (extensive) on-screen output generated by running conversion.sh. Perhaps the most useful output file is all.gc, which can be inspected directly if gc-info and Gmaj do not convey the desired information. The first line is a parenthesized specification of the species tree, with numbers 1, 2, … assigned to branches (so that the phylogenetic branch associated with each conversion event can be indicated). Subsequent lines contain detailed information for each paralogous pair of intervals where a conversion event was detected, using the following fields:

index : index for each pair of paralogous sequences
species : name of species
beg1 : start position of the first sequence (i.e., the first paralogous interval in the named species)
end1 : end position of the first sequence
beg2 : start position of the second sequence (i.e., the second paralogous interval)
end2 : end position of the second sequence
strand : strand of the second sequence
length : length of the first sequence
identity : fraction of identical nucleotides for the two sequences
GC_len : length of the gene conversion region
P-value : p-value for the conversion test
GC_beg1 : start position for the conversion region in the first sequence
GC_end1 : end position for the conversion region in the first sequence
GC_beg2 : start position for the conversion region in the second sequence
GC_end2 : end position for the conversion region in the second sequence
direction : direction of conversion, encoded as:

0. unknown
1. the first sequence is converted
2. the second sequence is converted

C1_name, C1_start, C1_end, C1_orient : ortholog of the first sequence in the outgroup species
C2_name, C2_start, C2_end, C2_orient : ortholog of the second sequence in the outgroup species
event_number : the event_index as given in the "all.remove_redundancy.gc" file
tree_branch : indication of where the conversion event occurred in the tree topology
ortholog_block1 : indices of blocks that contain the ortholog of the first sequence
ortholog_block2 : indices of blocks that contain the ortholog of the second sequence
ortholog_status : status of orthologs in the outgroup species, encoded as:

0. no orthologs;
1. triplet test; only paralog #1 has an ortholog
2. triplet test; only paralog #2 has an ortholog
3. quadruplet test; both paralogs have orthologs
4. the conversion event covers almost the entire duplicated region; we look for evidence that the duplication preceded the speciation