RepeatMasker Documentation

Interpreting Results | RepeatMasker uses| Sensitivity | Selectivity | Repeat databases | References | Changes

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). On average, almost 50% of a human genomic DNA sequence currently will be masked by the program. Sequence comparisons in RepeatMasker are performed by the program cross_match, an efficient implementation of the Smith-Waterman-Gotoh algorithm developed by Phil Green.

Input format:

Sequences can be pasted in or uploaded as files, both in fasta format. Multiple fasta format sequences may be pasted in at once or may be contained within a file. Fasta format looks like this:

```
>Sequence1
ACGTGCGCGATCGCCTGCTAGGCGTACGTCGCAG
GCGATCGATGTGCTAGATCAGATGACA
>Sequence2
GGGCTAGATTAGCACCACATACATCGCTCA
```

The submission form contains a text field for the full pathname of the file containing the sequence data on the local system (i.e. where the Netscape browser is running). By pressing the "Browse..." button, you can use a file selection box to select the file without having to type the path. When running the browser on a MacIntosh the browse button works but the file name can not be typed in. On both the PC and Mac the sequence file needs to be saved as 'text only'.

Size limitations

In principle, there is no limit to the length of the query sequence or size of the batch file. However, the most common error message obtained by users is due to timing out of the connection during the submission of long sequences. Furthermore, longer sequences (> 50kb) are queued (when necessary), whereas shorter sequences are handled instantly (see also "sensitivity and speed" below. When you routinely submit large sequences it may be better to run RepeatMasker locally.

Output / return format

The program returns three or four output files for each query. One contains the submitted sequence(s) in which all recognized interspersed or simple repeats have been masked. In the masked areas, each base is replaced with an N, so that the returned sequence is of the same length as the original. A table annotating the masked sequences as well as a table summarizing the repeat content of the query sequence will be returned to your screen. Optionally a file with alignments of the query with the matching repeats will be returned as well.

In the "html" return format (default when the browser runs on a Mac or PC) all output is returned to your screen in one file. In the "tar file" return format the masked sequence(s) and alignments can be saved as compressed files. The "links" return format returns links to these output files in a text format (they look bad on the browser, but are fine when saved to your computer).

Options

Show alignments
When checked, alignments are returned in a file (ending in .aln) or to the screen. Alignments are shown in order of appearance in the query sequence.

Do not mask simple.../Only mask simple...
Regions of low complexity, like simple tandem repeats, polypurine and AT-rich regions can lead to spurious matches in database searches. By default they are masked along with the interspersed repeats.
With the option "Do not mask simple..." only interspersed repeats are masked. This may, for example, be preferred when the masked sequence will be fed to a gene prediction program.
Alternatively, with the option "Only mask simple...", one can mask only these low complexity regions, e.g. when you are only interested to quickly locate polymorphic simple repeats in a sequence.

Only mask Alus
By checking this option, you limit the masking and annotation to (primate) Alu repeats. 7SL RNA (the ancestral sequence of Alus), SVA (which contains several Alu sequences and a fragment of LTR5) and LTR5 are masked as well. This option only works for primate DNA.

Mask with Xs...
When checked, the repeat sequences are replaced by Xs instead of Ns. This allows one to distinguish the masked areas from possibly existing ambiguous bases or other stretches of Ns in the original sequence. However, when running BLAST searches (and maybe other programs) Xs are deleted

out of the query and the returned BLAST matches will have position numbers not necessarily corresponding to that of the original sequence.

Fixed-width columns
Since April 1999 the column widths in the annotation table are adjusted to the maximum length of any string occurring in a column; this allows long sequence names to be spelled out completely. Previously a fixed column width table was returned, which can still be obtained by checking this option button.

Other options

You can type in less frequently used options in UNIX command line style, like:

```
-div 20 -inv -GC 45
```

which will cause the program to only annotate and mask repeats less than 20% diverged, return the alignments in the orientation of the repeat consensus sequences, and use matrices optimal for a 45% GC background nucleotide distribution.

With the option **-div** you can limit the masking and annotation to a subset of less diverged (younger) repeats by choosing a maximum divergence level of the repeat copy to its consensus sequence. This option may be used to limit the masking to those repeats that are either specific to primates or another mammalian order for use in subsequent comparison of orthologous mammalian loci. On average, interspersed repeats have diverged 18% in human (~35% in mouse) from their consensus since the mammalian orders separated, so typing '-div 18' in the advanced options box limits masking to most primate specific repeats. Note that this method is rather crude, mostly since the range of deterioration of repeats of the same age is wide; many shared repeats may go unmasked and vice versa..

Neutral mutation patterns differ significantly depending on the GC richness of a locus and we have calculated optimal scoring matrices for the alignment to consensus sequences in a range of background GC levels. Usually, RepeatMasker calculates the percentage of the sequence consisting of Gs and Cs and uses the appropriate matrices. However, the program defaults to using 'average' 43% GC matrices when the query is shorter than 2000 bp or a batch file is analyzed. Short sequences are less likely to share the GC level of the locus. For example, CpG islands and exons are more GC rich than the surrounding DNA, whereas a LINE1 element usually is more AT rich than the background. In a batch file, RepeatMasker analyses all sequences together with the same matrices. The percentage GC in all the sequences combined may be inappropriate for some sequence entries; using high GC level matrices in AT rich sequences (and vice versa) may result in false masking.
One can override this behavior in two ways:
With the option **-gc** you can set the GC level to a certain percentage; e.g. '-gc 37' lets the program use matrices appropriate for 37% GC background. This could be useful, for example, when you have a batch file of ESTs from a single locus with a known GC level.
Alternatively, the **-gccalc** option forces RepeatMasker to use the actual GC level of a short sequence or the average GC level of a batch of sequences. The latter sequences, for example, may be contigs or reads in a sequencing project.

RepeatMasker transparently fragments large sequences in fragments of 60 kb with 2 kb overlaps. The **-frag** option allows one to change the size of these fragments. Fragmentation was implemented to allow the size of sequences and sequence batches to be unlimited. It also can improve repeat detection when a genomic sequence contains regions of DNA with significantly different GC levels (isochores); sets of scoring matrices are chosen based on the GC level of a fragment. The only visible effect of the fragmentation is in the alignment files, where alignments at the edges of the fragments can be duplicated and/or truncated.

Alignments are shown in the orientation of the query sequence. The option **-inv** will return alignments in the orientation of the repeats.

In the process of finding all repeats, RepeatMasker temporarily cuts out most full-length elements, young LINE1 3' ends, and close to perfect simple repeats are deleted (both in human and rodent settings) to unearth any possible underlying older repeat in which these elements have inserted or expanded. The option **-nocut** skips the above deletion step in the default procedure. RepeatMasker is generally more sensitive including the deletion step.

When the option **-xsmall** is used a sequence is returned in the .masked file in which repeat regions are in lower case and non-repetitive regions are in capitals.

The option **-small** causes the whole masked sequence to be returned in lower case, with repeats replaced by 'x's (or 'x's if combined with -x).

DNA source

Interspersed repeats are specific to a (group of) species, dependent on the time of activity of the source transposable element. About half of the repeats identified in human DNA are specific to primates, i.e. they amplified after the eukaryotic radiation some 100 million years ago. Most repeats that can be identified in mouse DNA are specific to rodents, due to higher activity and faster mutation rates in the rodent lineage. RepeatMasker has separate protocols optimized for analysis of rodent and primate genomes. Interspersed repeats in other mammals have not been so well catalogued as yet. Among these, artiodactyl queries are treated best by RepeatMasker, but repeats specific to other orders are also present.

The numbers of different repeat consensus sequences against which queries of different species are compared gives an impression of how far the different libraries are developed:

	# of repeats	total bp
primates	563	664160
rodents	466	487006
other mammals	347	243730
other vertebrates	52	53994
Drosophila	65	167423
Arabidopsis	98	275516
grasses	27	67789

Note that the majority of sequences against which rodent and especially other mammalian queries are compared are repeats identified in the human genome and thought to predate the mammalian radiation.

Whereas the mammalian libraries represent heavily manipulated and expanded versions of Repbase libraries, the non-mammalian libraries were extracted with very limited curation. The vertebrate (chicken, Xenopus, etc.) and grasses (maize, rice) libraries are especially fetal. No summary tables are returned for these two.

Speed and sensitivity

On average, with default settings, a 10 kb human cosmid will be analyzed in about 30-40 seconds if no one else is using the server at the time.
For longer sequences the required time increases pretty much linearly with the sequence length. Sequences shorter than 10 kb are analyzed disproportionally faster. This is partially due to the program, e.g. a batch file of 200 human sequences of 400 bp (total 80 kb) is analyzed within 2 minutes, but we also have implemented a queuing system for sequences longer than 10 kb, making the request of lower priority the longer the query sequence. The speed is further somewhat dependent on the repeat content of the sequence; repeat dense regions, especially Alu-rich regions, are analyzed faster.

The program can be run at three levels of speed or sensitivity. The only difference between these settings is the minimum match or word length in the initial (not quite) hashing step of the cross_match program (see the cross_match/phrap documentation). The "slow" setting will take about 3 times longer and will find and mask 0-5% more repetitive DNA sequences than the default setting. The "quick" settings miss 5-10% of the sequences masked by default, but will be 3 to 6 times faster. The alignments may extend more or be somewhat more accurate in the more sensitive settings as well.

At the sensitive settings RepeatMasker currently finds, on average, 47% of human genomic DNA to be derived from interspersed repeats. RepeatMasker is very sensitive in comparison with other programs, although comparison to some is skewed because of the use of much smaller databases.

Selectivity and matches to coding sequences

The cutoff Smith-Waterman scores for masking interspersed repeats are conservative, since masking of one short potentially interesting region generally is more harmful than not masking a number of hard to find matches. If there are any false matches, they tend to have scores close to the cutoff, which is 225 for most repeats, 300 for the low-complexity LINE1 search, and 180 for the very old MIR, LINE2 and MER5 sequences.
We tested for the occurrence of false matches in randomized and in inverted (but not complemented) DNA. To check a variety of conditions, four 150 to 400 kb DNA fragments were analyzed ranging in GC level from 36% to 54%. To retain seeds for Smith Waterman alignments, randomization was done at the 10 bp word level. Note that the inverted sequences retain the low complexity and simple repeat patterns of the original sequences. Even at sensitive settings, for which false matches are most likely, this version of RepeatMasker reported no (false) matches at all to interspersed repeats in the randomized or inverted sequences. No simple repeats were reported in the randomized queries.

RepeatMasker returned only a single probably false match (71 bp) when analyzing a batch of 4440 coding regions in human mRNAs (7,200,000 bp) at sensitive settings. The coding regions were collected from GenBank, based on annotations, filtered for the presence of complete ORFs and initiator methionines, and made more or less non-redundant. When each coding region was analyzed individually using the -gccalc option, 5 matches (414 bp, 0.006%) were falsely masked (156 bp at default speed, 76 bp at quick settings). In this analysis each sequence was analyzed with matrices chosen based on the actual GC level, even for very short sequences, while in the batch analysis of the coding regions the 'average' 43% GC matrices were used.

Use in database searches

RepeatMasker is most commonly used to avoid spurious matches in database searches. Generally this step is strongly recommended before doing BLASTN or BLASTX equivalent searches with mammalian DNA sequence.

The most common concern is of course if RepeatMasker ever masks coding regions.
We found that false matches in coding regions are extremely rare, but did identify 38 genuine fragments of interspersed repeats (4214 bp) in the (annotated) coding regions of the 4440 human mRNAs (7.2 Mb) analyzed (excluding annotated coding sequences of LINE1 elements and endogenous retroviruses). We verified matches with lower scores by comparing the translation products to close homologous or redundant entries in the database (the repeat matching regions always were exactly missing). In the majority of these cases, the sequences appear to be improperly annotated or to represent either artificially or naturally defective mRNAs (e.g. alternatively spliced exons comprised of a small fragment of a repeat). Genuine overlaps of interspersed repeats with coding sequences usually involve terminal regions of the ORFs. Since the transposable element derived region is unique to the protein in that (group of) species, the masking does not interfere with database searches.

However, some cautionary comments are necessary. First, a few active cellular genes are derived from transposable elements. For example, I have identified 7 examples of human genes derived from (DNA transposon) transposases. These genes will be partially masked by a (related) DNA transposon in the repeat database. EST and cDNA matches beyond the masked region should alert you.

Also be aware that RepeatMasker screens for small RNA pseudogenes and will therefore mask the active small RNA genes as well (I think the tRNA list is complete, I stopped adding snRNAs unless I found an indication that they have created many pseudogenes). The number of matches to small RNAs are listed in the overview table; (close to) exact matches are possibly active genes, although related active genes not in the database may show diverged matches.

A final caution relates to the fact that 3' UTRs of transcripts are about as dense in interspersed repeats as intergenic regions are. Thus, many ESTs are completely masked as repetitive DNA. I recommend that, when you compare a genomic sequence against the EST database or use ESTs as a query in nucleotide searches, you search with the unmasked sequence as well; use a long minimum match (word length/ word size) like 40 bp to identify exact matches and avoid most background. Unfortunately the maximum word length that can be used in the NCBI BLASTN program is 18 (apparently due to memory limitations).

Use in association with gene prediction programs

Predicting genes from a masked sequence faces several problems. First, one should not mask low complexity regions, e.g. to avoid masking trinucleotide repeats in coding regions. But even with only interspersed repeats masked, gene prediction programs may fail to identify exons correctly. As mentioned above, sometimes tail ends of coding regions may have originated from transposable elements. Even if no coding regions have been masked, splice sites may be compromised; e.g. the polypyrimidine region that is part of the acceptor splice site may be contained within a repeat.

Thus, I generally recommend to run a gene prediction program on unmasked DNA (as well) and compare the predicted genes and exons with the RepeatMasker output. Some gene prediction program allow you to force certain exons out of the predictions (e.g. often the old ORFs of LINE1 elements and endogenous retroviruses are included in genes). Work is also in progress at several sites to incorporate RepeatMasker into gene prediction programs, in which cases matches to repeats are weighted in along with the other parameters used.

Other uses

Many people mask repeats before designing primers or oligo probes from sequence data. I've been told often that primers/probes designed from regions unmasked by RepeatMasker have a much better success rate. A cautionary note here is that unmasked regions not necessarily are unique in the genome (e.g. many lower copy repeats are not in the database yet) and experiments should be performed as if no filtering against repeats has been done.
The alignments can help in designing primers from sequences that are completely masked. Regions that diverge much from the consensus are less likely to misbehave than others.

RepeatMasker is sometimes used during assembly of large genomic sequences. This procedure probably is most useful in very Alu rich regions; in that situation I recommend to only mask the Alus, and maybe limit the masking to those Alus less than 15% diverged (-div 15).

How to read the results

The annotation file contains the cross_match output lines. It lists all best matches (above a set minimum score) between the query sequence and any of the sequences in the repeat database or with low complexity DNA. The term "best matches" reflects that a match is not shown if its domain is over 80% contained within the domain of a higher scoring match, where the "domain" of a match is the region in the query sequence that is defined by the alignment start and stop. These domains have been masked in the returned masked sequence file. In the output, matches are ordered by query name, and for each query by position of the start of the alignment.

Example:

```
 1306 15.6  6.2  0.0 HSU08988  6563  6781  (22462) C  MER7A    DNA/MER2_type    (0)   336   103
12204 10.0  2.4  1.8 HSU08988  6782  7714  (21529) C  TIGGER1  DNA/MER2_type    (0)  2418  1493
  279  3.0  0.0  0.0 HSU08988  7719  7751  (21492) +  (TTTTA)n Simple_repeat      1    33   (0)
 1765 13.4  6.5  1.8 HSU08988  7752  8022  (21221) C  AluSx    SINE/Alu        (23)   289     1
12204 10.0  2.4  1.8 HSU08988  8023  8694  (20549) C  TIGGER1  DNA/MER2_type  (925)  1493   827
 1984 11.1  0.3  0.7 HSU08988  8695  9000  (20243) C  AluSg    SINE/Alu         (5)   305     1
12204 10.0  2.4  1.8 HSU08988  9001  9695  (19548) C  TIGGER1  DNA/MER2_type (1591)   827     2
  711 21.2  1.4  0.0 HSU08988  9696  9816  (19427) C  MER7A    DNA/MER2_type  (224)   122     2
```

This is a sequence in which a Tigger1 DNA transposon has integrated into a MER7 DNA transposon copy. Subsequently two Alus integrated in the Tigger1 sequence. The simple repeat is derived from the poly A of the Alu element. The first line is interpreted like this:

```
  1306    = Smith-Waterman score of the match, usually complexity adjusted
        The SW scores are not always directly comparable. Sometimes
        the complexity adjustment has been turned off, and a variety of
        scoring-matrices are used.
  15.6    = % substitutions in matching region compared to the consensus
  6.2     = % of bases opposite a gap in the query sequence (deleted bp)
  0.0     = % of bases opposite a gap in the repeat consensus (inserted bp)
  HSU08988 = name of query sequence
  6563    = starting position of match in query sequence
  7714    = ending position of match in query sequence
  (22462) = no. of bases in query sequence past the ending position of match
  C       = match is with the Complement of the consensus sequence in the database
  MER7A   = name of the matching interspersed repeat
  DNA/MER2_type = the class of the repeat, in this case a DNA transposon
            fossil of the MER2 group (see below for list and references)
  (0)     = no. of bases in (complement of) the repeat consensus sequence
            prior to beginning of the match (so 0 means that the match extended
            all the way to the end of the repeat consensus sequence)
  2418    = starting position of match in database sequence (using top-strand numbering)
  1465    = ending position of match in database sequence
```

An asterisk (*) in the final column (no example shown) indicates that there is a higher-scoring match whose domain partly (<80%) includes the domain of this match.

Note that the SW score and divergence numbers for the three Tigger1 lines are identical. This is because the information is derived from a single alignment (the Alus were deleted from the query before the alignment with the Tigger element was performed). The program makes educated guesses about many fragments if they are derived from the same element (e.g. it knows that the MER7A fragments represent one insert). In a next version I can identify each element with a unique ID, if interest exists (this could help to represent repeats cleaner in graphic displays).

Alignments

Alignments are shown in order of appearance in the query sequence. These alignments may be most generally useful for designing PCR primers in a region full of repeats. It is possible to get primers that work in a whole genome, when the 3' end of it lies in a region of (even a common) repeat that is very different from the consensus. Alignments are shown in the orientation of the query sequence unless the option -inv is typed in in the option box.

Here is an example of an alignment of a MIR spanning an Alu element deleted in an earlier step:

```
665  28.45  2.93  5.02  g5129s420  7350  7882  (1924)  C  MIR#SINE/MIR  (1)  261  28  3

  g5129s420         7350 ATCATAACAAACATTTAT--GGTGCCTCCTATGGAGCAGGGATTTTGCTT 7397
                           v     v           i i  i v     viv    v i v v  v
C MIR#SINE/MIR       261 ATAATAACCAACATTTATTGAGCGCTTACTATGTGCCAGGCACTGTTCTA 212

  g5129s420         7398 AGGACTCTGAACTATAT---CTTACTT-GTCTTCATTAAAAACCTTATGA 7443
                           vi i iv   i        i i   i  i    i  v    i
C MIR#SINE/MIR       211 AGCGCTTTACA-TGTATTAACTCATTTAATCCTCA-CAACAACCCTATGA 164

  g5129s420         7444 AAAAGGTACTATTATTAACTGGGGXTGGGTTGTTTAACAGATAAGAAAGC 7787
                         iiv              v i         iii   v      i  i  i
C MIR#SINE/MIR       163 GGTAGGTACTATTATTATCC---------CCATTTTACAGATGAGGAAAC 123

  g5129s420         7788 TTAAGAATTAGAGAGATAAATTATCTTGCTTAAGGTAACACAGTTAACAA 7837
                          v i v  i      i v  v  v     ii     v      i  ii
C MIR#SINE/MIR       122 TGAGGCA-CAGAGAGGTTAAGTAACTTGCCCAAGGTCACACAGCTAGTAA 74

  g5129s420         7838 GCATTAG-GTCAAAGTTTGAACTCGGGCAGTCTGACTACAGAGCCC 7882
                          iivi    i iiii  i    i i         i  v     i
C MIR#SINE/MIR        73 GTGGCAGAGCCGGGATTCGAACCCAGGCAGTCTGGCTCCAGAGTCC 28

Transitions / transversions = 1.96 (45 / 23)
Gap_init rate = 0.03 (8 / 234), avg. gap size = 2.38 (19 / 8)
```

In cross_match alignments the mismatches are indicated, where "-" indicates an insertion/deletion, "i" a transition (G<->A, C<->T) and "v" a transversion (all other substitutions). The position of the deleted Alu in the query is indicated with an "X".
The lines in the annotation table describing this match appear as:

```
 665  28.4  2.9  5.0  g5129s420  7350  7467 (533) C  MIR   SINE/MIR  (1)   261  149
2222  10.2  2.7  0.0  g5129s420  7468  7762 (238) C  AluSg SINE/Alu  (7)   303    1
 665  28.4  2.9  5.0  g5129s420  7763  7882 (118) C  MIR   SINE/MIR  (113) 149   28
```

Discrepancies between alignments and annotation

Most discrepancies between alignments and annotation result from adjustments made to produce more legible annotation. This annotation also tends to be closer to the biological reality than the raw cross_match output. For example, adjustments often are necessary when a repeat is fragmented through deletions, insertions, or an inversion. Many subfamilies of repeats closely resemble each other, and when a repeat is fragmented these fragments can be assigned different subfamily names in the raw output. The program often can decide if fragments are derived from the same integrated transposable element and which subfamily name is appropriate (subsequently given to all fragments). This can result in discrepancies in the repeat name and matching positions in the consensus sequence (subfamily consensus sequences differ in length).

Some other discrepancies are specific to LINE elements. These repeats do not appear as complete elements in the consensus database. This is mostly a result of the contrast in conservation over the length of its sequence during its evolution in the mammalian genome; the ~3 kb ORF2 region of LINE1 has been very conserved, whereas the untranslated regions and ORF1 to a lesser degree have evolved very fast. Thus the 3' end or 5' end of an ancient LINE1 does not even remotely resemble that of the currently active LINE1, whereas the coding region for reverse transcriptase is closely related. Thus, many subfamilies have been defined for both the 5' and 3' UTRs (25 and 50, resp.) of LINE1 elements in human DNA, whereas only three ORF2 entries are present in the database. It is not only hard to extend all subfamilies from the beginning to the end, but it also appears that different 3' ends may have been associated with the same 3' ends, and vice versa. On top of that, including 50 full length (6.2-8 kb) LINE1 elements in the database would make the program very slow. LINE1 elements therefore are presented in the database in 3 (or more) pieces, and the program tries to put these pieces together as well as possible. As a result both the names of the repeats and position numbering in the consensus sequence are generally different in the alignments than in the output file. The LINE2 elements are likewise broken up in the databases, in 3' UTRs for different subfamilies and one ORF2 region.

The 3' UTR of LINE1 subfamilies ranges from 500 bp to over 2000 bp (in L1MC/D3), and the length of the 5' UTR is even more variable, even between subfamilies that show strong similarity in the 3' UTR. To allow the LINE1 fragments to be put together, all position numbers in older LINE1 subfamilies are adjusted to the position of ORF2 (the conserved part of LINE1) in a complete L1PA2 element. Since some older elements have much longer 5' UTRs or ORF1-ORF2 linker regions than L1PA2, this sometimes results in the assignment of negative position numbers for the 5' end of LINEs.

Finally, you may find large discrepancies in position numbering if an element includes tandem repeat units. For example, MER109 contains multiple

~300 bp repeat units; this can lead to overlapping matches. In the output such matches are fused.

The summary (.tbl) file

The summary file is pretty much self explanatory. Below is an example.

```
==================================================
file name: A-355G7.fasta
sequences:          1
total length:  139958 bp
GC level:       41.03 %
bases masked    91491 bp ( 65.37 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:              46       12182 bp     8.70 %
      ALUs          41       11603 bp     8.29 %
      MIRs           5         579 bp     0.41 %

LINEs:              42       52641 bp    37.61 %
      LINE1         38       52296 bp    37.37 %
      LINE2          4         345 bp     0.25 %

LTR elements:       20       13441 bp     9.60 %
      MaLRs         10        5618 bp     4.01 %
      Retrov.        4        5131 bp     3.67 %
      MER4_group     3        1439 bp     1.03 %

DNA elements:        8        1741 bp     1.24 %
      MER1_type      7        1114 bp     0.80 %
      MER2_type      1         627 bp     0.45 %
      Mariners       0           0 bp     0.00 %

Unclassified:        5        9215 bp     6.58 %

Total interspersed repeats:  89220 bp    63.75 %


Small RNA:           0           0 bp     0.00 %

Satellites:          0           0 bp     0.00 %
Simple repeats:     20        1647 bp     1.18 %
Low complexity:      9         437 bp     0.31 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element

The sequence(s) were assumed to be of primate origin.
RepeatMasker version  11/06/98               default
ProcessRepeats version  06/16/98
```

The four main classes mentioned in this table are well defined (see my 1996 review in COGD) and form a good basis for a summary or visual presentation of the repeats in a locus. Among the subclasses, some uncertainty of classification remains; it is especially hard to predict if an LTR is derived from an endogenous retrovirus or a non-autonomous LTR element. Also, not all subclasses are listed and the total for the classes is often higher than the sum of the sub classes. Note that the "MER" subclasses and the different MER interspersed repeats are not necessarily related to each other. The term MER (MEdium Reiterated repeats) was introduced for purely administrative purposes to give the beast a name. I named the MER1, MER2, and MER4 groups after the first member of each group that was identified as an interspersed repeat.

The program tries very hard to find out which repeat fragments were derived from the same insertion event of a transposable element. The estimated number of events still tend to be an overestimate.

The 'bases masked' number is calculated from the total number of Xs in the masked sequences (before these are changed to Ns or lower case letters). The other numbers are derived from the annotation (.out) file. Discrepancies between the 'bases masked' number and the sum of 'total interspersed repeats', small RNA, satellites and low complexity are generally very small. They are mostly accounted for by unmasked regions between flanking identical simple repeats, annotated as one stretch if fewer than 10 bases separate them, and fragments of repeats shorter than 10 bp which are not annotated but are masked.

Low-complexity DNA and simple repeats

By default, along with the interspersed repeats, RepeatMasker masks low-complexity DNA. Simple repeats (micro-satellites) can originate at any site in the genome, and therefore have an interspersed character. Other low-complexity DNA, primarily poly-purine/ poly-pyrimidine stretches, or regions of extremely high AT or GC content will result in spurious matches in some database searches as well (especially in the ungapped BLASTN searches). For example, extremely AT-rich regions consistently will give very low probability matches to mitochondrial DNA in BLASTN searches. The settings are very stringent, and we think that few if any sequences informative in database searches are masked as low-complexity DNA. However, one may opt to skip the low-complexity masking, for example when using RepeatMasker in conjunction with a gene prediction program.
Under the current settings a 100 bp stretch of DNA is masked when it is >87% AT or >89% GC, a 30 bp stretch has to contain 29 A/T (or GC)

nucleotides. The settings are slightly more stringent than the original settings, partly because the new, gapped BLAST programs are less sensitive to short regions of low complexity. In coding regions I did not find extensive regions (>10 bp) masked as low complexity DNA that would not be masked by the combined XNU and SEG filters routinely used in BLASTX.

Finding polymorphic simple repeats

Although RepeatMasker does a good job in masking simple repeats to avoid spurious matches in database searches, it is not written to find and indicate all possibly polymorphic simple repeat sequences. Only di- to pentameric and some hexameric repeats are scanned for and simple repeats shorter than 20 bp are ignored. Combining the "Only mask simple.." button option with a "div" option (e.g. -div 10) will produce a list of simple repeats that are 90% or more perfect. However, this list may not be not complete; e.g. two perfect 40 bp long (CA)n repeats interrupted by 10 Ts are aligned in one piece and may be reported as having > 10% divergence from the consensus. Of course most hexameric and longer unit repeats won't be reported either. A site dedicated to identifying polymorphic tandem repeats can be found at UTSW .

Reference repeat databases

The interspersed repeat databases screened by RepeatMasker are based on the repeat databases (Repbase Update) copyrighted by the Genetic Information Research Institute (G.I.R.I.). The Repbase Update database contains annotation of most repeats with respect to divergence level, affiliation, etc. The nomenclature of the interspersed repeats in the output of RepeatMasker is nearly identical to that of the reference database which in most cases corresponds to that in the literature.

Scoring matrices

We have calculated statistically optimal scoring matrices for the alignment of neutrally diverging (non-selected) sequences in human DNA to their original sequence. These matrices have been in use since the May 1998 release. The matrices were derived from alignments of DNA transposon fossils to their consensus sequences (Arian Smit, Arnie Kas & Phil Green, in preparation...). A series of different matrices are used dependent on the divergence level (14-25%) of the repeats and the background GC level (35-53%, neutral mutation patterns differ significantly in different isochores).

These matrices are (close to) optimal for human genomic sequences longer than 10 kb, for which length the GC level usually is representative of the isochore in which the sequence lives. However, the GC level of small fragments can diverge a lot from the surrounding (e.g. a fragment spanning a CpG island, a GC rich exon or an AT-rich LINE1 element) and RepeatMasker defaults to using matrices derived for a 43% GC background when a sequence is shorter than 2000 bp or when a batch file is submitted. When the appropriate background GC level is known, this can be entered with the -gc option.

Reference

We haven't published a paper on RepeatMasker yet, unless you call this expanding help file a publication. We'd appreciate it if you could refer to the web site in your publications (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org).

Literature

OVERVIEW
Smit, A.F.A. (1996) Origin of interspersed repeats in the human genome. Curr. Opin. Genet. Devel. 6 (6), 743-749.
Smit, A.F.A. (1996) Structure and evolution of mammalian interspersed repeats. PhD dissertation, USC. (lots of otherwise unpublished information here, available under order number 9636751 at the UMI web site)

SINE/Alu
Schmid, C. W. (1996). Alu: structure, origin, evolution, significance, and function of one-tenth of human DNA. Prog Nucleic Acids Res Mol Biol 53, 283-319.
Jurka, J. (1996) Origin and evolution of Alu repetitive elements. In " The impact of short interspersed elements (SINEs) on the host genome. Maraia, R.J., editor. Springer Verlag.
Batzer, M. A., Deininger, P. L., Hellmann Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M., Schmid, C. W., Zietkiewicz, E., and Zuckerkandl, E. (1996). Standardized nomenclature for Alu repeats. J Mol Evol 42, 3-6.

SINE/MIR & LINE/L2
Smit, A. F. A., and Riggs, A. D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res 23, 98-102.

LINE/L1
Smit, A. F. A., Toth, G., Riggs, A. D., Jurka, J., Ancestral mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246, 401-417.

LTR/MaLR
Smit, A. F. A. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res 21, 1863-72.

LTR/Retroviral

Wilkinson, D. A., Mager, D. L., and Leong, J. C. (1994). Endogenous Human Retroviruses. In The Retroviridae, J. A. Levy, ed. (New York: Plenum Press), pp. 465-535.

DNA/all types
Smit, A. F. A., and Riggs, A. D. (1996). Tiggers and other DNA transposon fossils in the human genome. Proc Natl Acad Sci USA 93, 1443-8.

Improvements and new features

June 1997

The database of human/mammalian-wide repeats was expanded 2.5 fold. Among the new additions are the (long) internal sequences of endogenous retroviruses.

Databases of repeats from other species than primates, rodents or artiodactyls can now be screened, although the program is not optimized to do so and the quality of the databases is not at the same level.

Through optimization of the cross_match searches, the program more sensitive and selective, especially with regard to detection of low complexity sequences and old LINE1 elements.

The RepeatMasker output is now processed by a second script to create annotation ready for database submission. Some of the more obvious improvements in the output are (i) overlapping matches are generally resolved, (ii) LINE1 fragments are annotated with position numbers as in a full L1 element, and (iii) when an Alu or LINE1 is fragmented information from both or all fragments is used to assign a subfamily name.

Alignments are shown without interruption by other cross_match output and in the order of appearance in the query sequence.

A summary table has been added which shows, among other things, the repeat composition of the query sequence.

September 1997

- major expansion of the rodent libraries and significant update of the human libraries as well, especially in LINE1 elements.
- scripts modified to accommodate new entries in databases
- simple repeats masking optimized by including pentamers and using a more stringent matrix
- several bugs fixed (e.g. sequences without repeats are now counted)
- table now displays the parameters used

June 1998

- the program is more robust and accepts most 'almost but not quite fasta' format files
- large sequences are analyzed in fragments of 100 kb to reduce the memory requirements of the program. Similarly files with very many sequence entries are divided up. You shouldn't notice any of this in the output files.
- matrices are used that are optimal for the divergence level of the repeats to which the query is compared and the background nucleotide composition.
- another big update of the human repeat databases.
- the small RNA sequences have been corrected and expanded (all tRNAs should be there now)
- the summary table now lists the amount of small RNA (pseudo)genes, simple repeats and low complexity DNA identified
- close to perfect simple repeats, full-length shorter interspersed repeats and young LINE1 3' ends are temporarily excised from the sequence (in both human and rodent analysis) to allow better detection of any underlying repeats.
- the "Skip simple, low complexity region masking" really skips all simple repeats now
- alignments are shown in the orientation of the query sequence
- among many bugs fixed is one involving sequence names including a number between parentheses

December 1998

This version uses the 1998 cross_match release. The difference for RepeatMasker is mainly in the complexity adjusted length of the matches that function as kernels for Smith Waterman alignments and the matrix dependent adjustment of the score for complexity of the alignment.

The full description ('>') lines are now retained in the masked file.

The .out file table is returned with flexible length columns allowing the full length of long query sequence names to be displayed. Optionally, the old fixed width table can still be obtained.

Simple repeat and satellite masking has been improved again; their annotation has changed a bit, most notably they are now all listed in the orientation of the query sequence

Several new options are available:

- A mRNA/EST option prevents false masking due to inappropriate matrix choice and low complexity matches to LINE1 elements in short GC rich regions like coding regions.
- You can limit the masking to Alus when masking primate DNA
- You can limit the masking to younger repeats by setting a maximum allowed divergence to the consensus sequence
- The sequences identified as repeats can be returned in lower case (rest in capitals) rather than masked out by Ns or Xs.
- You can set the background GC level (determining which matrices are used) overriding the program's calculations.

Among bugs fixed since May 1998 are those responsible for distorted output for sequences with names ending in .seq and for sequences without a header line. Also, sequence files from PCs and Mac with hidden carriage returns are handled appropriately.

April 1999

All the command line options are now available on the web site.

The default return format of the annotation file is changed, hopefully in a way that does not interfere with any type of parsing; the width of the columns is now adjusted to the longest entry in that column, allowing query names to be spelled out in full, and usually leading to narrower tables.

Arabidopsis, Drosophila, and grass repeat libraries were added; other repeat libraries were updated.

Three measures were taken to eliminate the (few) false positives:
- Use of the actual average GC level of sequences in a batch file may sometimes lead to false masking (or failure to mask) in sequences that diverge largely from the average. Thus, by default, all batch files are now analyzed with the innocuous 43% matrices.
- one entry, responsible for 90% of false masking in GC rich regions, is deleted from the 'tough L1' library.
- the matrix used for identification of the most diverged sequences in very GC rich regions, based on too little data and too much extrapolation, was 'too easy' on the mismatches and has been adjusted.

Thanks to these measures the 'mrna' option is not necessary and has been removed.

A bug is fixed that led to (wildly) improper annotation for some sequences fully consisting of repeats (all bases masked). A series of lesser bugs were taken care of. New bugs were introduced, probably.

For further information and to obtain a local copy go to the RepeatMasker Download Page.

Institute for Systems Biology
This server is made possible by funding from the National Human Genome Research Institute (NHGRI grant # RO1 HG002939).