BIO 4342 Lecture on Repeats

Jeremy Buhler

June 14, 2006

1 How RepeatMasker Works

Running RepeatMasker is the most common "first step" in annotating genomic DNA sequences. What exactly does it do?

- Given a database of known repeats, RepeatMasker finds copies of these repeats in a DNA sequence.
- We will consider several aspects of its implementation.
  - *Data source*
  - *Operation*
  - *Issues and Limitations*

2 Data Source: Repeat Libraries

How does RepeatMasker know what to look for?

- It uses *repeat libraries* containing a particular type of repeat (Alu), or several types of repeats from a particular species.
- Standard repeat libraries are supplied by the *RepBase* project (from the Genetic Information Research Institute).
- Repeat sequences in a library may be *consensus* sequences compiled from many divergent copies. In other words, the most frequent base among all copies is given for each position of the repeat.
- Repeat copies are aligned with multiple alignment tools, such as ClustalW, with substantial manual curation.

Why store only consensus sequences for repeats rather than the complete list of original sequences?

- The database of consensus sequences is much smaller, but this is not the only reason.
- Interspersed repeats mostly arise from transposons.
- There is/was some “active” version of the transposon, but most copies are inactive and so have decayed over time by neutral mutation.
- Consider two copies of a long repeat, A and B, diverged neutrally for equal lengths of time from a common consensus C.
- If the repeat is long relative to the number of mutated positions, then the mutated positions in A and B are unlikely to overlap.
- Hence, $\mathrm{dist}(A, B) \approx \mathrm{dist}(A, C) + \mathrm{dist}(B, C) \approx 2\mathrm{dist}(A, C)$.
- *Conclusion*: one copy of the repeat is likely to be closer in sequence to the consensus than it is to other copies of the same repeat.
- The computational cost of discovering conserved sequences goes up sharply as the distance between them increases, so scanning a sequence with the consensus is a *much* less expensive alternative to scanning with the individual repeat copies.

What kind of sequences are found in repeat libraries other than transposons?

- There are simple repeats, such as AGAGAGAG or ATCATCATC.
- There are non-repetitive but “low complexity” DNA composed primarily of one or two of the four possible nucleotides. For example, AAAGGAAGAAAAGAGAAAAAGAG.
- There are known micro- and mini-satellite sequences (e.g. centromeric and subtelomeric families).
- There are known small noncoding RNAs (tRNA, snRNA, scRNA, rRNA).
- There are vector/linker and *E. coli* sequences. These sequences often show up as *contamination* in eukaryotic DNA sequence. Vectors and linkers are used as part of the sequencing protocol, while *E. coli* is a common host organism for maintaining libraries of DNA fragments in plasmids.

3 Methods: Finding Repeats

How does RepeatMasker use a repeat library to discover known repeats in a sequence?

- The basic approach is BLAST-like: compare the sequence to the contents of the library, looking for significant similarity.
- "Classic" RepeatMasker used *CrossMatch* for comparing sequences. This is the same comparison tool used by Phrap to find fragment overlaps for assembly.
- Modern RepeatMasker also offers the option to use WU-BLAST. This *MaskerAid* option is about as sensitive as using CrossMatch, but it runs up to 10× faster.
- For commonly used libraries, RepeatMasker uses hand-tuned similarity search parameters (scores, gap penalties, score threshold (not E-value) for reporting matches.

Typically, a match to *part* of a repeat in a sequence is significant enough that RepeatMasker reports it, even if the rest of the repeat cannot be recognized.

- Why is this important?
- First, not all parts of a repeat may be equally conserved. For example, retroviral-like elements are often recognized by three specific regions, corresponding to the *gag*, *pol*, and *env* genes of the original retrovirus.
- We may not be able to see similarity to the repeat's consensus in poorly conserved regions.
- Second, some types of transposon relics, especially retroposons, tend to occur as partial copies of varying length. This arises because reverse transcription of new retroposon copies starts from one end but may not reach all the way to the other end. The partial retroposon is still inserted back into the genome.

There's one more subtlety to consider when finding repeats.

- Repeats frequently occur *inside* of other repeats!
- This arises historically whenever a newer repeat is inserted in the middle of an older one.

- RepeatMasker tries hard not to mistake the two parts of such an "interrupted" repeat for two distinct repeats.
- The exact algorithm is family-specific and not well documented, so I can't make any promises about how well it works.
- *Warning*: this specialized processing of split repeats is probably tuned to work best in mammals. No guarantees about, say, fruit flies.

4 Issues Arising When Using RepeatMasker

Issue: Repeats are hard to find.

- Neutral mutation of a sequence can cause frequent indels, as often as every 10 bases in old LINE repeats.
- BLAST-like similarity search algorithms use heuristics that are very sensitive to indels. A high indel frequency can cause many real alignments to be missed.
- RepeatMasker is slow in large part because it must change the default settings of CrossMatch or WU-BLAST to be more resistant to indels, at a significant cost to efficiency.
- Even so, highly diverged or fragmented repeats can still be missed.
- Even when repeats are correctly found, RepeatMasker may recognize them over their entire length. Hence, the ends of a repeat might be left unmasked.
- *Moral*: be very suspicious of "new" features found in sequence immediately adjacent to a masked repeat.

Issue: Use the right libraries!

- Lineage-specific repeats will not be found unless you specify right libraries.
- Normally, you just give correct sequence type (primate, rodent, artiodactyl, etc.), and RepeatMasker uses right set of libraries.
- In some cases (e.g. mosquito), you may have to specify libraries explicitly by name.

- Beware of using the wrong library – you may find repeats, but they will be labeled incorrectly!
- *Example*: human Alu and mouse B1/B2 are both tRNA-derived SINEs.
- Repeat-masking human with a mouse library incorrectly labels many Alus as B1/B2.

Issue: Not all repeats are best found by RepeatMasker.

- Short tandem repeats and low-complexity DNA can better be found by other algorithms.
  - **Dust** identifies low-complexity sequence.
  - **TRF** finds tandem repeats, both short *and* very long.
- Non-coding RNA sequences mutate in a way that preserves their structure, but not their sequence.
- To find RNA families, use a tool designed for the family of interest, or a general RNA search tool like Infernal.
- Repeats that occur in just a few copies or are unique to a particular sequence generally don't make it into the standard RepBase libraries.
- Hence, RepeatMasker will not find such repeats.
- **Examples**: chromosome-specific repeat families in human, tandem or other localized block duplications, inverted repeats

In conclusion, let's be careful out there!