Design and Use of RepeatMasker

Jeremy Buhler
HHMI / BIO4342
Tutorial Workshop

Parts of RepeatMasker

- **Programs**
  - Smit AFA, Hubley R, and Green P. "RepeatMasker-Open 3.0." 1996-2004. *http://www.repeatmasker.org*.
  - CrossMatch / WU-BLAST for comparisons
- **Data**
  - RepBase library
  - *http://www.girinst.org*

Overview

- Sources of repetitive sequence data
- How RepeatMasker finds repeats
- Issues and limitations

Data Source

- Uses a library of known repeat seqs
- Supplied by RepBase project
- Repeats in RepBase are carefully curated, typically by hand.

Page 265

Novosib-4_CR

Novosib-4_CR is an autonomous DNA transposon - a consensus.

Submitted: 00-0000		**Accepted:** 31-May-2006
Key Words: Novosib; DNA transposon; Interspersed Repeat; 8-bp TSD; transposase; Novosib superfamily; Novosib-4_CR		
Source: Chlamydomonas reinhardtii	**Organism:** Chlamydomonas reinhardtii	**Taxonomy:** Eukaryota; Viridiplantae; Chlorophyta; Chlorophyceae; Chlamydomonadales; Chlamydomonadaceae; Chlamydomonas
[1]	**Authors:** Kapitonov,V.V. and Jurka,J.	
	Title: Novosib-4_CR, a family of autonomous Novosib transposons from the green algae genome.	
	Journal: *Repbase Reports 6(5), 265-265 (2006)*	

Abstract:
Novosib-4_CR is a young family of autonomous transposons. The consensus sequence was derived from several copies that are ~98% identical to each other. These transposons are characterized by 8-bp target site duplications and 20-bp terminal inverted repeats.

Derived:
[1] (Consensus)

Download Sequence - Format:
IG, EMBL, FASTA

References:

An example summary report for a repeat family published by RepBase

Consensus Sequences

- A repeat family is usually summarized by a *consensus sequence*

```
accgataggtatacgtatca-tttacgatac
atcgct-ggtttacgcgtcaattcaggatgc
accggt-tgtttacgtagcaatctaggatac
```

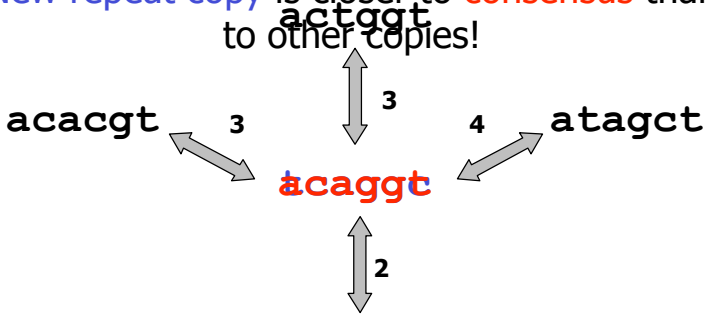
```
accgat-ggtttacgtatcaatttaggatac
```

Why Consensus Sequences?

- Faster to compare one sequence to genome than many
- Consensus can actually be better than individual instances for discovering new copies of a repeat.

Utility of Consensus

New repeat copy is closer to consensus than to other copies!



Types of Repeats in Library

- Interspersed (Alu, LINE, MIR, …)
- Simple (`agagagag`, `atcatcatc`, …)
- Micro- and mini-satellites
- Noncoding RNAs (tRNA, rRNA, snRNA, …)
- Common contaminants (E. coli, vectors)

Overview

- Sources of repetitive sequence data
- How RepeatMasker finds repeats
- Issues and limitations

The Basics

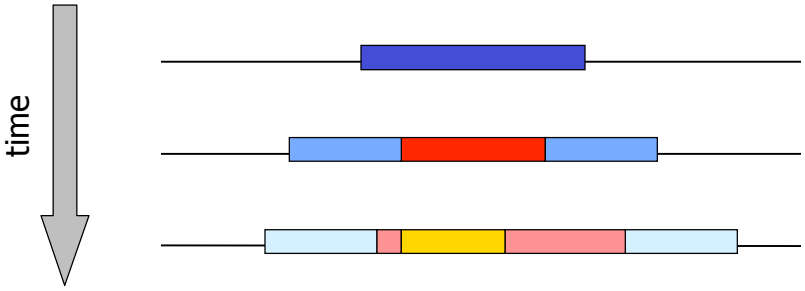
- Uses BLAST-like tool to compare libraries to query sequence
- Cross-Match (P. Green) – traditional
- WU-BLAST (W. Gish) – 10x faster!

Partial Repeats

- RepeatMasker will cheerfully report an incomplete match to a repeat.
- Detects best-conserved parts
- Some repeats (retroposons) typically incomplete

Nested Repeats

- RepeatMasker tries to detect nesting



- (Please don't ask me how)

Overview

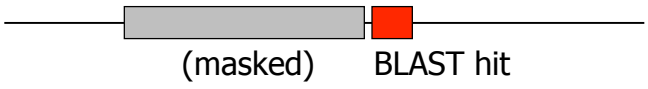
- Sources of repetitive sequence data
- How RepeatMasker finds repeats
- Issues and limitations

Library Choice

- Make sure to use correct libraries for your target species
- (Commonly used organisms have preselected library lists)
- Danger: mis-identifications!

Incomplete Masking

- Highly diverged repeats can be tough to find
- Might leave ends of a repeat unmasked



- Is this really a new feature?

Use the Right Tool

- Tandem repeats and duplications
  - Dust (short)
  - TRF (long)
- RNA
  - tRNAScan, Infernal, …
- Low-copy (chr-specific, inverted, …)
  - BLAST?

In conclusion…

Hey, let's be careful out there!