Repeats sequences

Hidden Markov Models
Repeats, Satellites & Transposable Elements
Transposable elements: junk dealers

Barbara McClintock

“Out standing in her field”

Transposable elements “jumping genes” lead to rapid germline variation.
Excision of transposon may leave a “scar”.

TR=tandem repeat
IR=inverted repeat

trucciformal structure

repaired DNA with copied TR and added IR
Millions of years of accumulated TE “scars”

Some genomes contain a large accumulation of transposon scars.
Estimated Transposable element-associated DNA content in selected genomes

- **H. sapiens**: 35%
- **Z. mays**: >50%
- **Drosophila**: 15%
- **Arabidopsis**: 2%
- **C. elegans**: 1.8%
- **S. cerevisiae**: 3.1%

TEs

Everything else
fun with bioinformatics jargon

ACRONYMS for satellites and transposons

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR</td>
<td>Short Sequence Repeat</td>
</tr>
<tr>
<td>STR</td>
<td>Short Tandem Repeat</td>
</tr>
<tr>
<td>VNTR</td>
<td>Variable Number Tandem Repeat</td>
</tr>
<tr>
<td>LTR</td>
<td>Long Terminal Repeat</td>
</tr>
<tr>
<td>LINE</td>
<td>Long Interspersed Nuclear Element</td>
</tr>
<tr>
<td>SINE</td>
<td>Short Interspersed Nuclear Element</td>
</tr>
<tr>
<td>MITE</td>
<td>Miniature Inverted repeat Transposable Element (class III TE)</td>
</tr>
<tr>
<td>TE</td>
<td>Transposable Element</td>
</tr>
<tr>
<td>IS</td>
<td>Insertion Sequence</td>
</tr>
<tr>
<td>IR</td>
<td>Inverted Repeat</td>
</tr>
<tr>
<td>RT</td>
<td>Reverse Transcriptase</td>
</tr>
<tr>
<td>TPase</td>
<td>Transposase</td>
</tr>
</tbody>
</table>

Class I TE, uses RT.

Class II TE, uses TPase.

Class III TE, MITEs*

Alu 11% of primate genome (SINE)

LINE1 14.6% of human genome

Tn7, Tn3, Tn10, Mu, IS50 transposons or transposable bacteriophage

retroposon=retrotransposon

*Class III are now merged with Class II TEs.
Types of repeat sequences

**Satellites** -- 1000+ bp in *heterochromatin*: centromeres, telomeres

Simple Sequence Repeats (SSRs), in *euchromatin*:

- **Minisatellites** -- ~15bp (VNTR)
- **Microsatellites** -- 2-6 bp

heterochromatin=compact, light bands
euchromatin=loose, dark bands.
microsatellite

a microsatellite in a dog (*canis familiaris*) gene.
This 8bp tandem repeat has a consensus sequence AGGATTTTT, but is almost never a perfect match to the consensus.
How do you recognize a repeat sequence?

- High scoring self-alignments
- High dot plot density
- Compositional bias

A repeat region in a dot plot.
Is there an evolutionary advantage of repeat sequences?

Repeat sequences are prone to

(1) locally: errors in replication

(2) non-locally: homologous recombination

Errors in replication (polymerase slippage) can lead to a change in the reading frame, eliminating a STOP codon, adding one, or translating to a different sequence entirely.

Neisseriae Gonorrhoeae evades the human immune system by randomly, periodically (weeks) changing the reading frame of the pilin surface antigen protein, aided by a 4-base microrepeat.
(How) do you align repeat sequences?

A: Don’t align. Mask them out instead.

B: Dynamic Programming with special EVD. Align just like any other sequence, but using a special null model to assess the significance of the alignment score. Use EVD to fit random scores.

Remember: Low complexity sequences will have high-scoring alignments randomly. For example:

```
ATTTATATAATTAATATATAAATATAATAAATATAAT
```

aligned to

```
TATTATATATATATATATATTATATTATATATATATATA
```

Random score is likely to have >50% identity!
Ariana Smit, Phil Green

Annotation Results

Compares your sequence to a *curated library of known repeats* to a query sequence: Returns: (1) **Location** and **type** of each repeat, and/or
(2) Query sequence with repeats masked (set to “N”)

<table>
<thead>
<tr>
<th>SW perc</th>
<th>perc</th>
<th>perc</th>
<th>query position in repeat</th>
<th>score div. del. ins. sequence</th>
<th>class/family</th>
<th>begin</th>
<th>end</th>
<th>(left)</th>
<th>repeat ID</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>194</td>
<td>10.5</td>
<td>2.6</td>
<td>0.0 chr1</td>
<td></td>
<td>Low_complexity</td>
<td>3</td>
<td>41</td>
<td>(0)</td>
<td>624</td>
<td>0</td>
</tr>
<tr>
<td>238</td>
<td>26.4</td>
<td>0.7</td>
<td>0.7 chr1</td>
<td></td>
<td>Simple_repeat</td>
<td>1</td>
<td>145</td>
<td>(0)</td>
<td>625</td>
<td>0</td>
</tr>
<tr>
<td>298</td>
<td>29.0</td>
<td>2.1</td>
<td>0.0 chr1</td>
<td></td>
<td>Simple_repeat</td>
<td>3</td>
<td>97</td>
<td>(0)</td>
<td>626</td>
<td>0</td>
</tr>
<tr>
<td>255</td>
<td>23.1</td>
<td>1.8</td>
<td>1.8 chr1</td>
<td></td>
<td>Simple_repeat</td>
<td>1</td>
<td>163</td>
<td>(0)</td>
<td>627</td>
<td>0</td>
</tr>
<tr>
<td>1864</td>
<td>13.8</td>
<td>0.0</td>
<td>0.7 chr1</td>
<td></td>
<td>SINE/Alu</td>
<td>5</td>
<td>287</td>
<td>(25)</td>
<td>628</td>
<td>0</td>
</tr>
</tbody>
</table>

www.repeatmasker.org
What is a stochastic emitter?
Markov models

A Markov “chain” is a network of “states” (circles) connected by “transitions” (arrows)

The probability of the next Markov state depends only on the current state.

Each state has a different “state identifier”: H=helix, E=beta sheet, L=loop
Markov model
Exercise: MMs for tongue-twisters.

how much wood would a wood chuck chuck if a wood chuck would chuck wood?

can you can a can as a canner can can a can?

I wish to wish the wish you wish to wish, but if you wish the wish the witch wishes, I won’t wish the wish you wish to wish.
The marble bag represents a probability distribution (PD) of amino acids, \( b \). (A PD of amino acids is a profile)

Each state has a unique “state identifier” and a PD (or multiple PDs)
Hidden Markov Model
HMMs consist of...

- Non-emitting source states (begin)
- Emitting states
- Non-emitting connector states
- Non-emitting sink states (end)
How to make a HMM from Blast results

<table>
<thead>
<tr>
<th>Score</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequences producing significant alignments:</td>
<td>(bits) Value</td>
</tr>
<tr>
<td>gi</td>
<td>18977279</td>
</tr>
<tr>
<td>gi</td>
<td>14521217</td>
</tr>
<tr>
<td>gi</td>
<td>14591052</td>
</tr>
<tr>
<td>gi</td>
<td>18313751</td>
</tr>
<tr>
<td>gi</td>
<td>729396</td>
</tr>
<tr>
<td>gi</td>
<td>1361925</td>
</tr>
<tr>
<td>gi</td>
<td>18312680</td>
</tr>
</tbody>
</table>

**QUERY**

| 3   | GI   |
| 259 | 58   |
| 54  | 53   |
| 275 | 274  |
| 269 | 268  |
| 272 | 271  |
| 505 | 504  |

First run BLAST. Output to MSA.
Assign parent sequence. Assign Insertion and Deletion states according to where gaps appear in the MSA.
Steps in making a profile HMM from a MSA

1. Choose a sequence to be the parent sequence, s (defines number of positions)

2. Read each sequence in MSA, i
   2.1. Create an Insert state if seq i has an insertion relative to seq s.
   2.2. Create a Delete state if seq i has a deletion relative to seq s.

3. At this point every position in every sequence is assigned to one Markov state.

4. Sum the emission profile for each state

5. Sum the transition probabilities.
Picking a parent sequence
(does it matter?)

- The parent defines the number of Match states
- A Match state should conserve the *chemical nature of the sidechain* as much as possible.
- A Match state implies *structural similarity*.

Which sequence is the best parent?
Filling in state probabilities

M=match state, I=insert state, D=delete state, MM=match-to-match, MI=match-to-insertion, MD=match-to-deletion, IM=insertion-to-deletion, Iself=I-to-I non-advancing, II=insertion-to-insertion, DM=deletion-to-insertion, DD=deletion-to-deletion.

Each element in the matrix is a probability, either an emission probability (M state) or a transition probability.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>E</th>
<th>G</th>
<th>K</th>
<th>I</th>
<th>G</th>
<th>K</th>
<th>V</th>
<th>K</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| A | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

Parent arrows:

HMM 26
Choose parent sequence ==> parent
Assign match states ==> m, to non-gap columns ==> c, of parent.
m = 0
For each column, c
   If (parent(c) == ".") { \ c is an insert state
      II(m) = II(m) + Count(.NOT.gap@c+1) \ next col is match
      ID(m) = Count(gap@m+1) \ next char is gap
      IM(m) = Count(.NOT.gap@m+1) \ next char is not gap
   } else { \ c is match state m
      m++ \ next match state
      for each amino acid, aa {
         profile(m,aa) = Count(aa@c) } \ use unweighted counts
      MD(m) = Count(gap@m+1) \ gaps at next match state
      MI(m) = parent(c+1) == "."? Count(.NOT.gap@c+1)| 0
             \ if next col is not a match state, count insertions
      MM(m) = Count(.NOT.gap@m+1) \ non-gaps at next match state
      DD(m) = Count(gap@m+1) \ gaps at next match state
      DM(m) = Count(.NOT.gap@m+1) \ non-gaps at next match state
   }
Normalize(MD+MI+MM)=1,(DD+DM)=1, (II+ID+IM)=1.
   \ convert counts to probabilities
Exercise 10: make a profile HMM

AGF----PDG
AGGYL--PDG
AG------PNG
SGFFLIPNG
SGF----EPNG

• Use seq 1 as parent.
• Draw non-zero Match, Delete, Insert states and transitions.
• Assign probabilities to all non-zero transitions.
• Draw a Logo for the Match state profile.
Added information

With Profile HMMs we allow *insertions* and *deletions* to have position-specific probabilities, and insertions have lengths.
Many uses of HMMs

Weather prediction
Ecosystem modeling
Brain activity
Language structure
Econometrics
etc etc

HMMs can be applied to any dataset that can be represented as strings.

The expert input is the “topology”, or how the states are connected.
Homolog detection using a library of profile HMMs

Get $P(S|\lambda)$ for each $\lambda$

1

MYSEQUENCE

$\rightarrow$

$\rightarrow$

$\rightarrow$

Pick the model with the max $P$

1.

$\rightarrow$

$\rightarrow$

2.

$\rightarrow$

$\rightarrow$

3.

$\rightarrow$

$\rightarrow$

4.

$\rightarrow$

$\rightarrow$

$P(s|\lambda_1)$

$P(s|\lambda_2)$

$P(s|\lambda_3)$

$P(s|\lambda_4)$
PFAM: Profile HMM libraries made by HMMER