Estimating Evolutionary Parameters for Protein Low Complexity Regions using an Approximate Bayesian Computation

Alexander Turco

November 27, 2022

McMaster University
SCIENCE
Department of Biology

Overview

- Background information
- Research Questions/Explorations
- Experimental Approach
- Results
- Conclusion and Future Work

What are LCR's?

Saccharomyces cerevisiae SRP40 Protein LCRs

>CAA82171.1(25-125) complexity=0.92 (15/1.90/2.20)
sssssssssssssssssssssssgessssssssssssssdssdssdsessssssssss
ssssssdsesssesdssssgsssssssssdesssesesede

>CAA82171.1(149-282) complexity=1.33 (15/1.90/2.20)
essssesssssgsssssesesgsesdsdsssssssssdsesdsesdsqsssssssssdsss
dsdsssdsssdsdsssssssssssdsdsdsdsssdsdssgssdsssssdsssdestssds
sdsdsdsdsgssse

>CAA82171.1(298-316) complexity=2.18 (15/1.90/2.20)
tpassnestpsasssssan

Shannon's Entropy - MAYBE

$$H = -L \sum p_i log_2(p_i)$$

LCR's Present in Unique Ways

Homorepeats

Consecutive iterations of a single residue

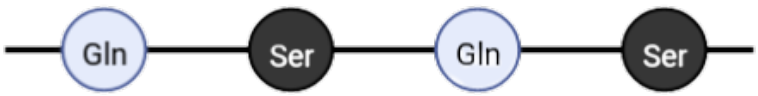
LCR's Present in Unique Ways

Homorepeats

Consecutive iterations of a single residue

Direpeats

Consecutive iterations of two ordered, different residues



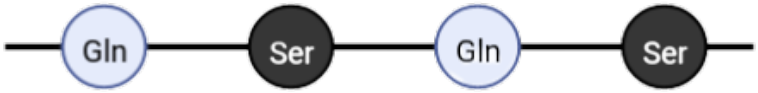
LCR's Present in Unique Ways

Homorepeats

Consecutive iterations of a single residue

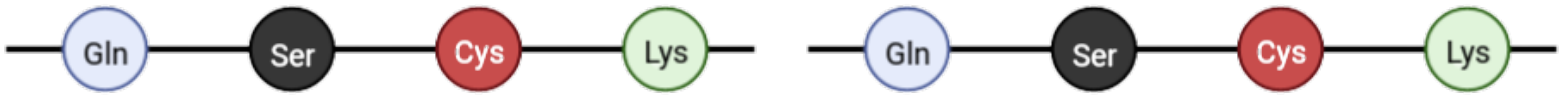
Direpeats

Consecutive iterations of two ordered, different residues

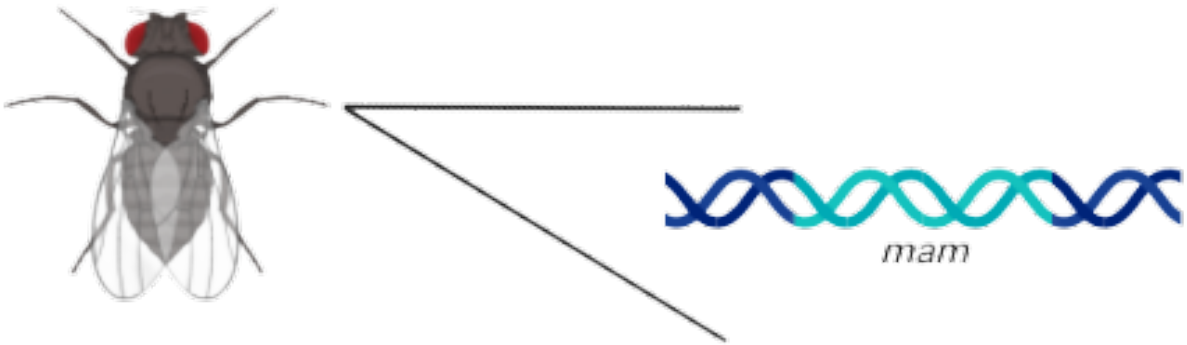


Tandem Repeats

Sequence of residues which are repeated a number of times



LCR's are Hypermutable

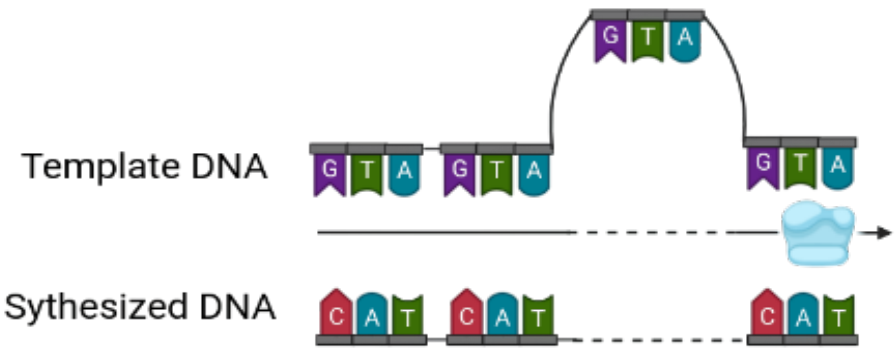


mam domain	Size (bp)	Amino Acid Substitutions	Amino Acid/ Total Substitutions
Unique	933	26	0.15
Repetitive	810	47	0.42

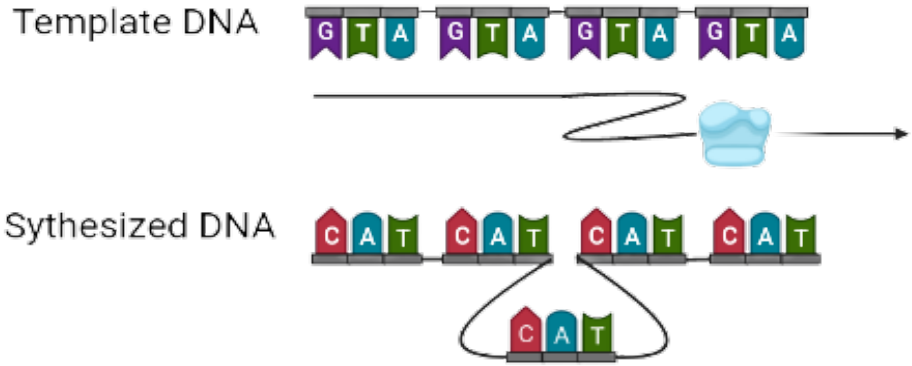
[1] Newfeld, Smoller, and Yedvobnick, 1991

Proposed Mechanisms of LCR Evolution

1. Polymerase Slippage/Slipped Strand Mispairing



Polymerase Slips Forward

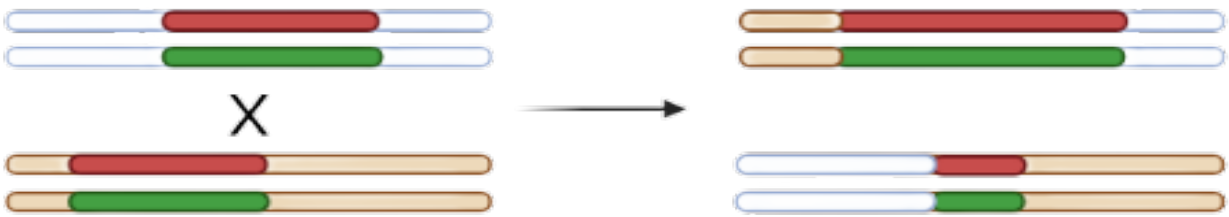


Polymerase Slips Backwards

[2] Levinson and Gutman, 1987; Sehn, 2015

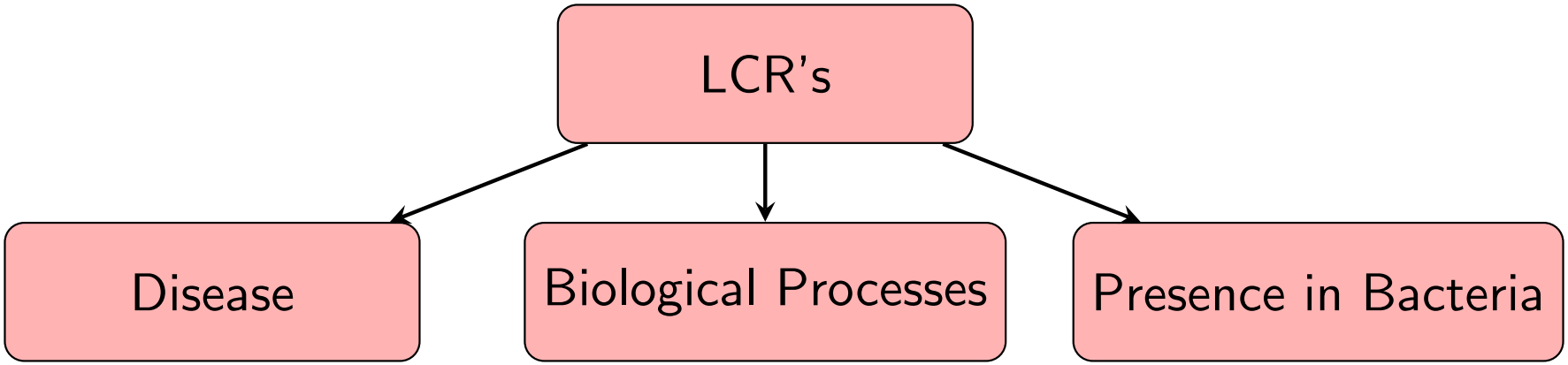
Proposed Mechanisms of LCR Evolution

2. Unequal Recombination

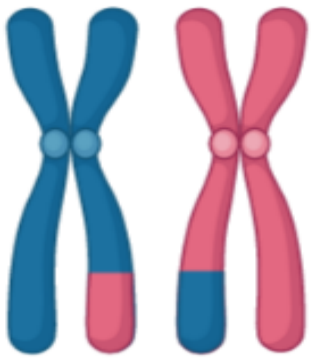


[3] Mirkin, 2007

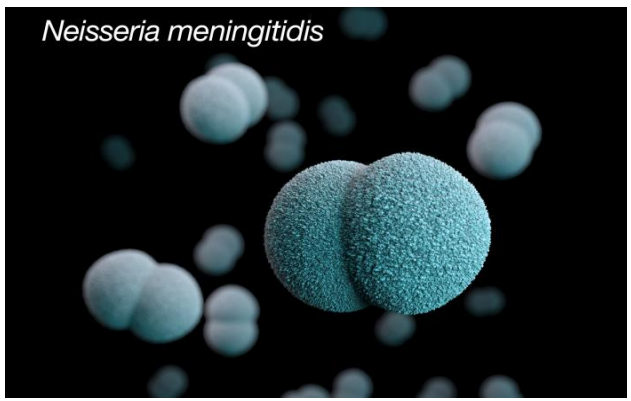
Why Care about LCRs and Their Evolution?



Huntington's Disease



Genetic Recombination



Neisseria meningitidis

What will this Study Explore?

- Estimation of evolutionary parameters (mutation rate, indel rates)
- Various models of insertions and deletions
- Summary statistics which best explain data

What Approach will be Taken?

Bayesian Statistics: Model-based statistical inference

$$p(D|\theta) \tag{1}$$

Likelihood

What Approach will be Taken?

Bayesian Statistics: Model-based statistical inference

$$p(D|\theta) \tag{1}$$

Likelihood

$$p(\theta|D) \tag{2}$$

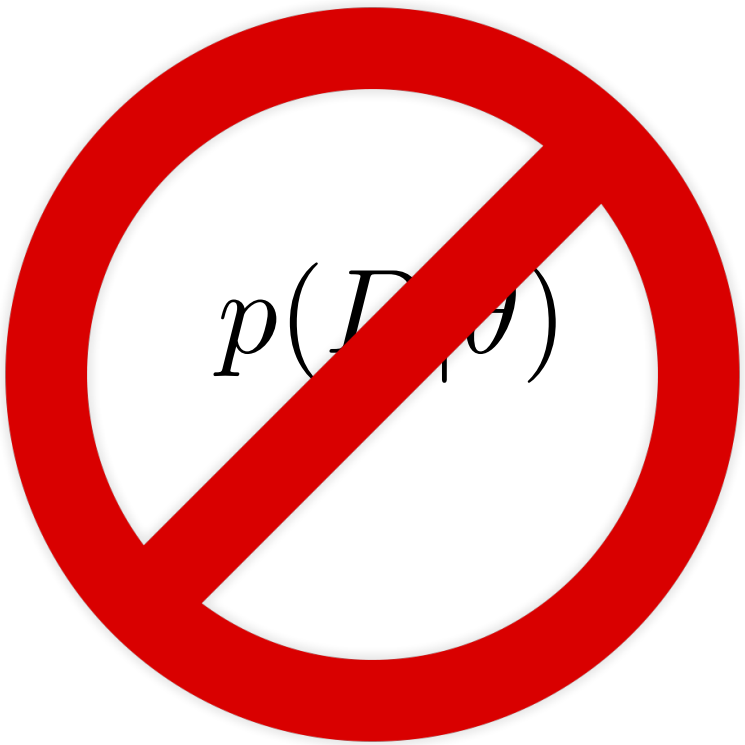
Posterior

Why use an ABC-MCMC

- The increasing complexity and magnitude of available data can make the likelihood difficult to calculate

$$p(D|\theta)$$

Why use an ABC-MCMC



- Calculation of the likelihood is replaced with a simulation step

MCMC for ABC

1. Propose a move from θ to θ' according to a transition kernel $q(\theta, \theta')$.
2. Generate simulated dataset D' using θ' and calculate S'.
3. If $\rho(S', S) \leq \epsilon$ continue to 4, otherwise remain at θ and go to 1.
4. Calculate

$$\alpha(\theta, \theta') = min(1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')})$$

5. Accept θ' with probability α, otherwise stay at θ.
6. Return to 1.

[3] Marjoram et al., 2003

MCMC for ABC: Modified Algorithm

1. Propose a move from θ to θ' according to the normal distribution

MCMC for ABC: Modified Algorithm

1. Propose a move from θ to θ' according to the normal distribution
2. Create a random protein sequence and mutate over many generations using θ' to generate simulated Dataset D'

MCMC for ABC: Modified Algorithm

1. Propose a move from θ to θ' according to the normal distribution
2. Create a random protein sequence and mutate over many generations using θ' to generate simulated Dataset D'
3. Calculate summary statistics for simulated dataset D'

MCMC for ABC: Modified Algorithm

1. Propose a move from θ to θ' according to the normal distribution
2. Create a random protein sequence and mutate over many generations using θ' to generate simulated Dataset D'
3. Calculate summary statistics for simulated dataset D'
4. If $d(S', S) \leq \epsilon$, go to next step, otherwise stay at θ and return to 1

MCMC for ABC: Modified Algorithm

1. Propose a move from θ to θ' according to the normal distribution
2. Create a random protein sequence and mutate over many generations using θ' to generate simulated Dataset D'
3. Calculate summary statistics for simulated dataset D'
4. If $d(S', S) \leq \epsilon$, go to next step, otherwise stay at θ and return to 1
5. Accept θ' with probability ?? Ask Brian ab this again lol

MCMC for ABC: Modified Algorithm

1. Propose a move from θ to θ' according to the normal distribution
2. Create a random protein sequence and mutate over many generations using θ' to generate simulated Dataset D'
3. Calculate summary statistics for simulated dataset D'
4. If $d(S', S) \leq \epsilon$, go to next step, otherwise stay at θ and return to 1
5. Accept θ' with probability ?? Ask Brian ab this again lol
6. Return to step 1

Simulation Step - MAYBE

Results

Conclusion/Future Work