

CREATION OF A LOW COMPLEXITY REGION EVOLUTION
SIMULATOR FOR USE IN AN APPROXIMATE BAYESIAN
COMPUTATION

ALEXANDER TURCO

April 19, 2023

¹ Department of Biology, McMaster University, Hamilton, ON, Canada

Contents

Abstract	4
Literature Review/Proposal	5
What are Low Complexity Regions?	5
Characteristics and Types of LCRs	5
Why care about LCRs?	6
How do LCRs Evolve?	7
What is an Approximate Bayesian Computation Markov chain Monte Carlo algorithm?	8
How will we use an ABC-MCMC - (Oct 28 Proposal, To See New Process Refer to Materials and Methods)	9
Materials and Methods (mid-year stuff Jan 20)	10
ABC-MCMC: The Algorithm	10
Parameters and Summary Statistics	11
Simulation Step: Creation and Mutation of Protein Sequences	12
Normalization and Euclidean Distance Calculation	12
One Sample T-Test for Probability of Acceptance	13
Results	14
Assessing the Evolution Simulator	14
Applying the Simulator to an ABC-MCMC	17
Discussion	18
Mutations Have the Ability to Destroy LCR Formation	18

Does the Length of a Repeat Play a Role in Insertions/Deletions	19
Why Parameters Could not be Estimated	19
Conclusion	20
References	21
Appendix 1: Low Complexity Region Evolution Simulator	24
Appendix 2: ABC-MCMC in C++	32
Appendix 3: Simulated Protein Summary Statistics	39
Appendix 4: Euclidean Distance and Normlization of Vectors	41
Appendix 5: One Sample t-test	43

Abstract

It has been shown that among eukaryotic proteomes, the most commonly shared peptide sequences tend to lack stable three-dimensional structures and contain low information and entropy content. These sequences, now known as low-complexity regions, lack diversity in amino acid composition, but recent research has demonstrated an important role played by these regions in a variety of cellular functions. Model-based analysis of amino acid sequences is a common approach utilized to study molecular evolution, however, due to the increasing complexity of available data, many model-based approaches have become intractable due to difficulties in calculating the likelihood function. To overcome the issue of intractability, this study employs an approximate Bayesian computation Markov chain Monte Carlo algorithm (ABC-MCMC), which utilizes a simulation step in place of calculating the likelihood function. Through the simulation of mutated protein sequences, this study aims to investigate the evolution of protein low-complexity regions by estimating parameters such as mutation and insertion/deletion rates. This work provides insight into the formation of low-complexity regions, and demonstrates the potential of using approximate Bayesian computation methods for studying evolutionary genetics.

Literature Review/Proposal

What are Low Complexity Regions?

For decades, it was believed that peptide sequences which lack the ability to form stable three-dimensional structures also lack specific biological function (Haerty and Golding 2010b). Interestingly, among eukaryotic proteomes, the most commonly shared peptide sequences are found to be sequences with a low information content which lack a stable three-dimensional structure (Marcotte et al. 1999; Bannen et al. 2007; Haerty and Golding 2010b). These sequences have been termed ‘low-complexity regions’ (LCRs) due to their low information content and entropy, as well as their lack of diversity in amino acid composition (Wootton and Federhen 1993; Coletta et al. 2010). LCRs are found in DNA as well as protein sequences and can present in a variety of ways, all of which skew the composition of amino acids in a different manner (Wootton and Federhen 1993; Mier et al. 2020). Homorepeats, direpeats, tandem repeats, and imperfect repeats are common definitions of LCRs based on the periodicity of amino acids in a given sequence, but not every LCR is defined by a specific pattern (Mier et al. 2020). Most of the time, these patterns are found to occur in non-coding regions and evolve with minimal selective pressure (Kruglyak et al. 2000). Further research is being done in order to uncover the function of LCRs in protein coding regions, as well as the evolutionary background of these repetitive regions (Huntley and Golding 2006). To investigate the process of LCR evolution, this study proposes an Approximate Bayesian Computation (ABC) approach which will enable the prediction of evolutionary parameters such as mutation rates and insertion/deletion rates.

Characteristics and Types of LCRs

Algorithms to detect the presence of low complexity regions in a sequence are available, and continue to be improved with further research into LCRs. Wootton and Federhen (1993) first developed an algorithm called SEG to find low complexity regions in protein sequences using information content (Huntley and Golding 2002). Information content is a common characteristic used to identify low complexity regions and in order to calculate the amount of information within a segment, the SEG algorithm implements Shannon’s entropy (Wootton and Federhen 1993; Battistuzzi et al. 2016). Shannon’s entropy (Shannon 1948) has been commonly used as a measure of complexity of a string of characters (Wootton and Federhen 1993; Coletta et al. 2010; Battistuzzi et al. 2016). This study will use the SEG algorithm and therefore Shannon’s entropy to assess the complexity of protein sequences. The less complex a sequence is (low variety of residues), the lower the entropy/information content of the sequence. Although LCRs are defined by their low information content, these regions have also been found to be hypermutable, and it is thought that throughout evolutionary history, they frequently gained and lost repeats (Kruglyak et al. 1998; Marcotte et al. 1999). In studying the *Drosophila melanogaster* gene mastermind, which encodes a highly repetitive nuclear protein, Newfeld et al. (1991) identified different patterns of evolutionary change between regions of high and low complexity. Repetitive regions were found to have a much higher rate of amino acid replacement, therefore the rate of evolution within these regions is higher than outside (Newfeld et al. 1991; Huntley and Golding 2000).

LCRs all share an overall low diversity of residues but present in unique ways as periodic or aperiodic repeats, which take on the form of homopolymers and heteropolymers (Wootton and Federhen 1993; Battistuzzi et al. 2016). Homopolymers/homorepeats are consecutive iterations of a single amino acid residue, and heteropolymers (direpeats, tandem repeats) are consecutive iterations of more than one residue that can be found in a variety of different patterns based on periodicity (Battistuzzi et al. 2016; Mier et al. 2020). Microsatellites, one of the best studied types of LCRs, commonly describe regions composed of tandem repeats that are typically made from anything between one to six nucleotides (Ellegren 2004). Although microsatellites normally refer to DNA sequences, it has been found that LCRs in proteins are comparable to microsatellites (DePristo et al. 2006). The molecular mechanisms involved in the process of evolution including slippage and unequal recombination are important for microsatellites and therefore protein LCRs as well (DePristo et al. 2006). A class of proteins which are related to, but slightly differ from LCRs are intrinsically disordered proteins (IDPs). IDPs are unable to form stable three dimensional structures and are characterized by low sequence complexity, biased amino acid composition, and high proportions of charged and hydrophilic residues (Wright and Dyson 2015). IDPs are composed of intrinsically disordered regions which are not necessarily defined by a low information content as LCRs are (Dunker et al. 2002; Haerty and Golding 2010b). In this study, we propose a focus on low complexity regions.

Why care about LCRs?

Proteins continue to be a large area of research due to their involvement in vital cellular processes and many human diseases. In protein sequence databases such as Swiss-Prot, the increase in the number of sequences and organisms represented has subsequently led to a decrease in the proportion of proteins containing LCRs (Coletta et al. 2010). On top of this, there is a lack of representation of LCRs in the protein data bank (Huntley and Golding 2002). Despite this underrepresentation of LCRs, they are known to be associated with several human neurodegenerative diseases and are thought to serve important biological functions (Huntley and Golding 2006; Coletta et al. 2010). It has also been found that the proteins of *Plasmodium falciparum* (the human malaria parasite) contain a high incidence of LCRs which has further highlighted the importance of both the evolution and function of LCRs (Gardner et al. 2002; DePristo et al. 2006). LCRs can appear as trinucleotide repeats which form repeated units of three nucleotides and are a well known form of deleterious mutation in humans (Ross et al. 1993). These are found to be the cause of diseases including fragile X syndrome, myotonic dystrophy, spinal atrophy, muscular atrophy, and Huntington's disease (Ross et al. 1993). These diseases can be broadly classified into two distinct groups, translated polyglutamine triplet repeat diseases and untranslated triplet repeat diseases (Everett and Wood 2004). Polyglutamine triplet repeat diseases, such as Huntington's disease, result in the formation of protein aggregates in the cell and occur due to expanded repeats being translated into expanded polyglutamine residues (Everett and Wood 2004). Untranslated triplet repeat diseases such as myotonic dystrophy and fragile X syndrome differ from polyglutamine repeat diseases as they contain trinucleotide repeats which are not translated into expansion within a mutant protein (Everett and Wood 2004).

LCRs have been associated with important biological processes such as genetic recombination, antigen diversification,

and protein-protein interactions (Karlin et al. 2002; Verstrepen et al. 2005; Kumari et al. 2015). The repetitive regions are thought to drive recombination events which alter genes and result in phenotypic variation (Verstrepen et al. 2005). In the genomes of *Haemophilus influenzae* and *Neisseria meningitidis*, LCRs are abundant and cause phase variation which gives the bacteria the ability to change their adherence patterns to host cells (Bayliss et al. 2001). This ultimately increases the fitness of the population and allows the bacteria to evade the host response (Bayliss et al. 2001). It was previously believed, based on structural evidence, that these hypermutable LCRs did not form stable structures but instead existed as solvent-exposed disordered coils (Wootton and Federhen 1993; Huntley and Golding 2002; DePristo et al. 2006). Using proteins from a non-redundant Protein Data Bank (PDB) dataset, Kumari et al. (2015) analyzed secondary structure content and surface accessibility and discovered that LCRs can form secondary structures within proteins. More specifically, in a large majority of identified LCRs, the analysis revealed the presence of more than one secondary structure, indicating that LCRs are found in regions where structure transition occurs (Kumari et al. 2015). Although more work is necessary to further understand the functions of LCRs, their role in genetic recombination, protein structure and function, and antigen diversity, highlight the importance of LCR research.

How do LCRs Evolve?

Although research surrounding the evolution of LCRs is lacking, there are two proposed mechanisms of microsatellite evolution, which can be applied to many forms of LCRs. The first, polymerase slippage or slipped strand mispairing, involves loops being formed in either the coding or template strand, which causes a misalignment of strands and results in either the insertion or deletion of repetitive motifs (Levinson and Gutman 1987; Ellegren 2004). It is believed that slipped strand mispairing is the predominant mode of mutation of LCRs, specifically in homopolymer sequences (Levinson and Gutman 1987). The second mechanism, unequal recombination, occurs when repetitive regions in homologous chromosomes do not align properly during meiosis, which results in the repetitive region being expanded in one chromosome and contracted in the other (Warren et al. 1997; Mirkin 2007). In order to gain more insight into the evolutionary background of a variety of organisms, researchers have created models of events such as slippage in order to estimate mutation rate and other evolutionary parameters (Kruglyak et al. 2000). In a study of 10,844 parent/child allele transfers at nine short tandem repeat loci, Brinkmann et al. (1998) discovered 23 mutations, all of which were either gains or losses of repeats. Of the 23 mutations, 22 were due to single repeat mutations, which is why it has been common to use the stepwise mutation model of Ohta and Kimura (1973), that assumes repetitive regions expand or contract by 1 unit at a specific mutation rate (Brinkmann et al. 1998; Kruglyak et al. 2000). There are however, major drawbacks of the stepwise mutation model including that lengths can become negative, and the collection of repeat lengths in a sample will not have a stationary distribution (Kruglyak et al. 2000). It was thought that more complex models of LCR evolution were necessary to gain more accurate results, thus Kruglyak et al. (1998) proposed a model that incorporated length dependent slippage events. This differed from the stepwise mutation model in that the balance between slippage events and point mutations produced an equilibrium distribution of repeats (Kruglyak et al. 1998).

Models of LCR formation including replication slippage support the historical belief that LCRs evolve neutrally. More recently, there has been increasing evidence suggesting that LCRs are also acted upon by selective pressure (Haerty and Golding 2010b). Kimura (1983), proposed the neutral theory of molecular evolution which suggests that selection does not play a role in the genetic diversity within and between species, rather genetic diversity is neutral (Nevo 2001). Evidence for the neutral evolution of LCRs relies on a large number of factors including both their lack of stable structure and function (Dunker et al. 2002; Haerty and Golding 2010b), and ability to frequently gain or lose repeats through replication slippage (Kruglyak et al. 1998; Marcotte et al. 1999; Huntley and Golding 2000). Support for a selective model of LCR evolution comes from the non-random patterns of changes within LCRs, the deleterious effect of their expansion in humans (Karlin et al. 2002), and their enrichment in proteins involved in transcription, DNA, protein binding, reproduction, and development (Huntley and Clark 2007; Haerty and Golding 2010a; Battistuzzi et al. 2016). In a study of orthologous mouse and human genes, Mularoni et al. (2007) found a significant negative correlation between repeat number and gene nonsynonymous substitution rate, indicating that proteins acted upon by strong selective pressure contain a large number of repeats conserved between the two species (Mularoni et al. 2007). Interestingly, the study also reported a significant positive correlation between repeat size difference and protein nonsynonymous substitution rate, demonstrating that events such as slippage and substitutions occur in proteins which undergo neutral evolution (Mularoni et al. 2007). It was later revealed in a study by Battistuzzi et al. (2016) in which 11 representative Apicomplexa genomes were analyzed, that neutral mechanisms were found to act on highly repetitive LCRs (homopolymers) whereas selective pressures were influenced by the heterogeneity and length of the LCR (Battistuzzi et al. 2016). This work only begins to unravel the complexities of the evolutionary patterns associated with LCRs.

What is an Approximate Bayesian Computation Markov chain Monte Carlo algorithm?

When studying molecular evolution, a common practice is to use model-based analyses of sets of DNA and amino acid sequences (Laurin-Lemay et al. 2022). This approach allows for the estimation of evolutionary genetic parameters such as mutation rates and insertion/deletion rates (Wu and Rodrigo 2015). Model-based statistical inference generally revolves around calculating the likelihood function, which represents the probability of the observed data under a chosen model (Sunnåker et al. 2013). The likelihood function therefore quantifies how well the data supports both the parameter values as well as the model (Sunnåker et al. 2013). However, due to an increase in the complexity and magnitude of available data, many current model-based analyses have become intractable by virtue of the likelihood function being difficult to calculate (Marjoram 2013). Approximate Bayesian computation (ABC) methods are rooted in Bayesian statistics and have been gaining popularity in areas such as genetics, as they bypass the calculation of the likelihood function (Sunnåker et al. 2013). The way in which they do this is by utilizing a simulation step in place of the calculation as a way to provide an estimate of the likelihood function (Marjoram 2013). Since there are many ways to approach a simulation, there are many different forms of ABCs. The more popular forms include ABC rejection methods, ABC Markov chain Monte Carlo methods (ABC-MCMC), and Sequential Monte Carlo ABC methods (ABC-SMC) (Marjoram 2013). This study proposes the use of an ABC-MCMC algorithm in order to estimate evolutionary parameters such as mutation and indel rates, and provide insight into the formation and evolution of protein LCRs.

The reason for proposing an ABC-MCMC in this study stems from the lack of a pre-existing model which explains how insertions and deletions work. Insertions and deletions alter the landscape of a sequence, making the likelihood calculation extremely challenging. Marjoram et al. (2003) originally proposed the algorithm for a MCMC method without the use of likelihoods. The algorithm first starts from a selected parameter value and proposes a move to a new parameter value based on a proposal distribution (Marjoram et al. 2003). Using this new parameter value, a dataset is then simulated and summary statistics are calculated, which makes it possible to quantitatively compare differences between the simulated dataset and the observed dataset (Marjoram et al. 2003). If the difference between summary statistics is small, the Hastings Ratio is calculated and the proposed parameter value can be accepted with a certain probability, then a new value is proposed and the process begins again (Marjoram et al. 2003). On the other hand if the difference in summary statistics between the observed and simulated data is very large, we propose a new parameter value and begin the algorithm again (Marjoram et al. 2003; Marjoram 2013). The use of this algorithm has enabled the analysis of complex problems which tend to arise in the areas of population genetics, ecology, epidemiology, and systems biology (Sunnåker et al. 2013). The group of Liepe et al. (2010) have been leaders in the use of ABCs for inference of genetic networks. This is evident through the creation of a software package they created called ABC SysBio which can implement ABC algorithms in a straightforward manner (Liepe et al. 2010; Marjoram 2013). Prior to the year 2000, there were essentially no papers published on ABCs (Marjoram 2013). As we enter into an era where larger and more complex data can be collected, the need for improved models is necessary, hence the large increase over the last decade in papers which mention ABC methods (Marjoram 2013).

How will we use an ABC-MCMC - (Oct 28 Proposal, To See New Process Refer to [Materials and Methods](#))

Using the algorithm mentioned above for an ABC-MCMC, this study aims to better understand the evolutionary background/formation of protein LCRs. An ABC-MCMC will enable the prediction of two important evolutionary parameters, mutation rate and indel rate. There is possibility for the estimation of other parameters which will be explored upon investigating the first two. We will utilize amino acid sequences in this study, one being the SRP40 protein found in *Saccharomyces cerevisiae*, which is extremely biased in composition. This protein sequence will act as our observed data and we will compare this observed data to our simulated data.

In terms of a simulation, we will use C++ to first generate a random amino acid sequence of a certain length. This randomly generated sequence will then be mutated over a number of generations in a two-step process. The first process is to choose a random Poisson deviate with a mean that is equal to the mutation rate multiplied by the total number of sites. A Poisson distribution is used here because mutation is a rare event and rare events can be modelled using this distribution. The value of the Poisson deviate yields the total number of sites in the simulated sequence which should be mutated at random. The second mutation process deals with amino acid expansion and in this case we iterate through each residue in the simulated protein sequence and scan for repeats. If a residue is part of a repeat, we take the total length of the repeat, multiply it by the

mutation rate and use this value as the mean of a random exponential deviate. We use the exponential distribution as it models waiting times between events. Based on the random exponential deviates assigned, we select the lowest value which represents the residue that will change fastest, and we alter that residue to either delete or insert a repeat at that position.

Once we simulate a protein sequence for a number of generations under certain parameter values, we need to obtain a set of summary statistics and compare the summary statistics of the observed and simulated data. We have proposed summary statistics based off notable characteristics such as protein length, number of LCRs, and the average entropy of the LCRs. There is the possibility for additional summary statistic characteristics upon exploration of the initially proposed characteristics. To quantitatively compare the differences between the observed and simulated data, we propose using a distance measure between the two vectors of summary statistics. This distance is just the norm of the vector observed-simulated. Along with this, we also propose the use of a threshold as a way to assess how close the two datasets are. If the distance between the two vectors of summary statistics is larger than this threshold, we can not accept the proposed parameter value and the algorithm begins again. On the other hand, if the distance is very small, we can move forward in the algorithm and potentially accept the new parameter value.

We intend to run the simulation under the same parameters many times and take the average of the produced summary statistic vectors before calculating the distance between observed and simulated data. It is also worth noting that each time we begin the algorithm again, new parameter proposals will be selected using random normal deviates. We hope to see the distance between summary statistics being minimized upon every initial iteration of the algorithm (burn-in), and then wander around according to the distribution that was utilized.

Materials and Methods (mid-year stuff Jan 20)

Custom scripts and commands utilized in this analysis can be found on GitHub at <https://github.com/opticrom/abcmcmc-thesis4c12>.

ABC-MCMC: The Algorithm

This study utilized the ABC Markov chain Monte Carlo algorithm, originally proposed by Marjoram et al. (2003). The algorithm begins from a randomly selected parameter value and follows the steps below.

1. If now at θ , propose a move to θ' according to a proposal distribution $q(\theta, \theta')$.
2. Simulate a dataset, D' using θ' .
3. If $D' \approx D$ proceed to step 4; else, output θ and return to 1.

4. Calculate the Hastings Ratio.
5. Accept, and output, the new θ' with probability h . Else return to, and output, θ , Go back to 1.

A custom C++ script was written to iterate through the algorithm for a desired number of simulations. The normal distribution was used to control how new parameter values were proposed. We simulated a dataset under the newly proposed parameter value, which consisted of a randomly generated protein sequence. The simulated protein was compared to a protein of known low complexity called SRP40, which is found in the model organism *Saccharomyces cerevisiae*. The protein sequence for SRP40 was obtained from the NCBI database. To quantitatively compare similarities between simulated and observed protein sequences, vectors of summary statistics (characteristics that describe the sequences) were created for each sequence.

The Euclidean distance between the observed and simulated protein vectors was calculated in order to determine if the newly proposed parameter value could be accepted. If the Euclidean distance between the SRP40 vector and the vector produced using the newly proposed parameter values was smaller than the distance between the SRP40 vector and the distance calculated using the current parameter values, the newly proposed parameter values were accepted. If this distance was larger, a one sample t-test was employed to determine the probability of accepting the newly proposed parameter values.

Parameters and Summary Statistics

Two parameters were estimated in this study, mutation rate and insertion/deletion (indel) rate. Mutation rate referred to the rate at which a single amino acid in a protein sequence changed into a different amino acid. Indel rate referred to the rate at which an amino acid was deleted or inserted from a protein sequence. For the purpose of this study, to determine if the length of a repetitive region played a role in insertions, any amino acid that was inserted into the simulated protein sequence was the same unit as the previous amino acid in the sequence.

Summary statistics were utilized to capture important information about simulated and observed protein sequences in order to assess how similar the sequences were. Three summary statistics were used which included, the length of the protein sequence, the number of LCRs in the sequence, and the average entropy of the LCRs. To identify LCRs and their corresponding entropies, the *Seg* algorithm was implemented (Wootton and Federhen 1993). The following *Seg* parameters were utilized to search for LCRs in proteins; a window length of 15, a trigger segment complexity of 1.9, and an extension segment complexity of 2.2. We selected these due to previous research which demonstrated that these parameter values would better detect regions of low complexity in eukaryotes which are typically longer and contain more repetitive repeats (Huntley and Golding 2000). Summary statistics were stored in vectors and normalized in order to prevent large values (for example the length of the protein sequence) from dominating when calculating the Euclidean distance.

Simulation Step: Creation and Mutation of Protein Sequences

To bypass calculation of the likelihood function, a custom C++ script was written to simulate the generation and mutation of protein sequences over numerous generations. Two random proteins of desired length were generated first, one that would be mutated based on the current simulation parameters, and one mutated based on the proposed simulation parameters. We generated two protein sequences 400 amino acids in length, similar to the length of the SRP40 protein in *S. cerevisiae*. This simulated protein sequence was then mutated in the following ways.

We iterated over the simulated protein sequence and assigned exponential deviates to each amino acid. We utilized the exponential distribution because it is commonly used to model waiting times between events. Mutation and indel rates served to act as the scale parameter (β), or mean of the distribution, indicative of the mean time until mutation occurs. In the case of mutation rate, the same rate was utilized across all sites, with the assumption made that repeats do not play a role in point mutation. In the case of the indel rate, we scanned for repeats, and if found, we multiplied the length of the repetitive segment by the indel rate, and utilized this new value as the scale parameter (β) for the exponential distribution.

Exponential deviates were stored in two vectors, one for deviates generated using mutation rate, and the other for deviates generated using the indel rate. We then identified a single deviate with the lowest value (based on both vectors), which represented the residue that mutated quickest. Depending on which vector the lowest deviate came from, we either altered the corresponding amino acid or inserted/deleted a repetitive amino acid. Upon mutating, deleting, or inserting an amino acid, we scanned the sequence to see if the mutation altered the landscape (interrupted a repeat, created a repeat), and subsequently generated new deviates for the amino acids that were affected.

We ran the simulation ten times for each newly proposed parameter value and took the average of all ten vectors of summary statistics prior to calculating the distance. On each simulation, we mutated the protein sequence for 50 generations before obtaining summary statistics. We plan on testing the program in the future with various numbers of iterations.

Normalization and Euclidean Distance Calculation

To determine the similarity between simulated and observed protein sequences, the distance between the vectors of summary statistics was calculated by employing a Euclidean distance measure (4). Before calculating this distance, vectors of summary statistics were normalized to prevent large values (such as protein length) from dominating the distance measure.

In order to normalize all elements of the vector to be between 0 and 1, maximum and minimum values for each summary statistic were required. For the length of the simulated protein, an upper limit of 1.5x the length of the SRP40 protein, and a lower limit of 0.5x the length of the SRP40 protein were set. Due to the length of the SRP40 protein being 406 amino acids long, the upper limit was equal to 609 amino acids, and the lower limit was equal to 203 amino acids (1). In terms of the

number of LCRs, the minimum value was set to 0 (no LCRs) and the maximum value was chosen based off the shortest LCR length. We selected 5 amino acids to be the length of the shortest LCR, resulting in the maximum number of LCRs being 81.2 (406/5) (2). Finally, for the average entropy of the LCRs, the minimum value was set to 0, and the maximum value was set to 4.3 (maximum entropy for amino acids). Normalization calculations are shown below.

$$normalized_length = \frac{(simulated_protein_length - SRP40_length + min_length)}{SRP40_length} \quad (1)$$

$$normalized_number_LCRs = \frac{number_LCRs}{max_number_LCRs} \quad (2)$$

$$normalized_average_entropy = \frac{average_entropy}{max_entropy} \quad (3)$$

$$d(P1, P2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \quad (4)$$

One Sample T-Test for Probability of Acceptance

A one sample t-test was employed in order to assess the probability of accepting newly proposed parameter values in cases where the Euclidean distance between vectors produced using the newly proposed parameters was larger than the distance between vectors produced using the current parameter values. Essentially the t-test provided a way to assess if the newly proposed parameter values are samples of data from the same distribution. For each set of newly proposed parameter values, 10 vectors of summary statistics were produced. From these 10 vectors, 10 distances were found, and the mean and standard deviation were calculated and subsequently used in the calculation of the t-statistic (5). Vectors of t-statistics and associated probabilities were stored, and the calculated t-statistic was used to estimate a p-value by linear interpolation.

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \quad (5)$$

Results

Assessing the Evolution Simulator

To be able to utilize the evolution simulator within an approximate bayesian computation, tests had to be done to ensure it was working properly. The first step in this study involved running simulations of LCR evolution for various mutation and indel rates. Simulations were run for 1000 iterations each, and on each iteration, the simulator would return the average entropy of the LCRs in the sequence, the total number of LCRs in the sequence, and the overall sequence entropy. To test the effects of each individual rate parameter, only 1 rate was adjusted while the other remained a fixed constant. Figure 1 and Figure 2 are the result of holding the mutation rate constant at 1, and only changing the insertion/deletion rate. It is expected that a low indel rate will lead to the formation of less LCRs and hence a higher sequence entropy. The results in Figure 1 match this expectation, showing the formation of only one to two low complexity regions that are quickly removed by mutations, and a relatively unaffected maximal protein sequence entropy.

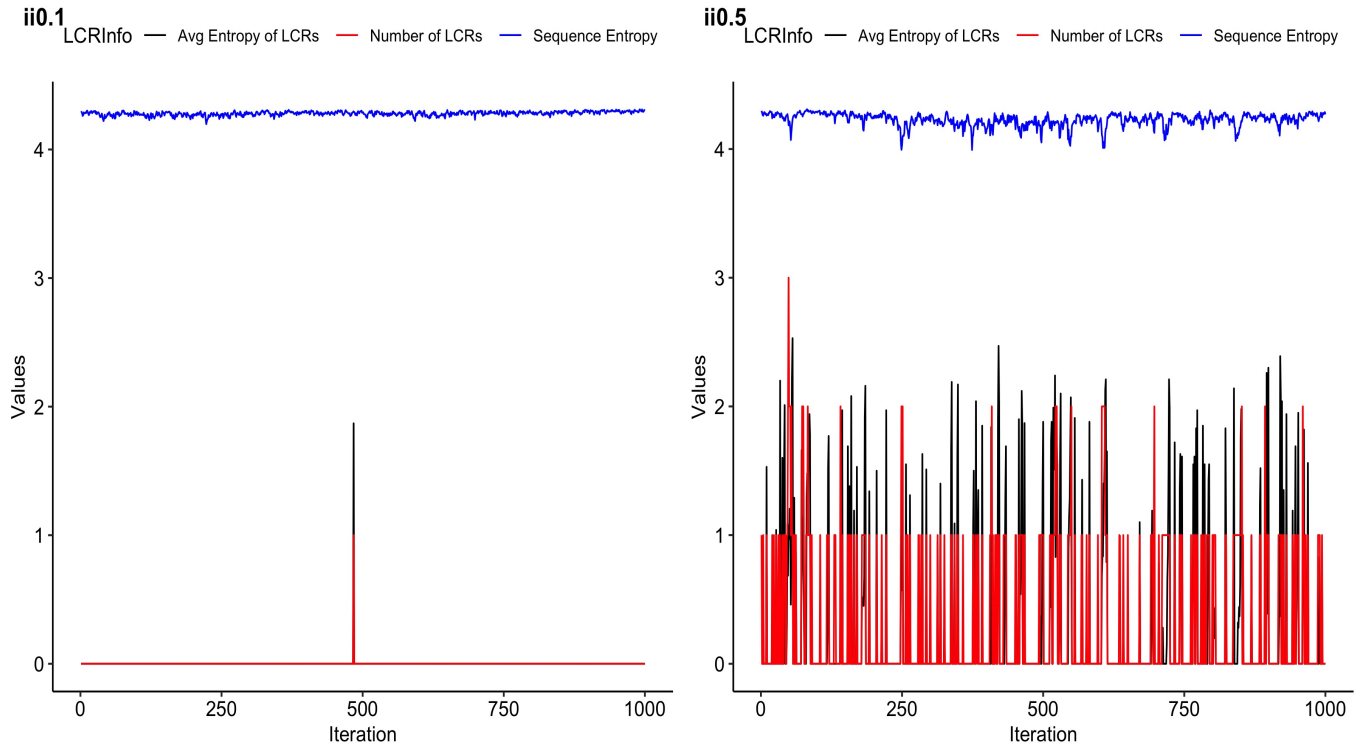


Figure 1: Entropy and number of LCR comparisons in a randomly generated protein sequence that has undergone simulated evolution. (Left) An indel rate of 0.1 was utilized. (RIGHT) An indel rate of 0.5 was utilized. A mutation rate of 1 was utilized for both.

As the indel rate is increased, it is expected that there should be more low complexity regions and hence, a lower average sequence entropy. Results in Figure 2 show the formation of many more LCRs that remain in the protein sequence, possibly because mutations don't have as much of a chance swamp out the formation of low complexity regions. These results also show

much more of a fluctuation in average sequence entropy compared to Figure 1, with decreases in entropy occurring where new LCRs are being formed.

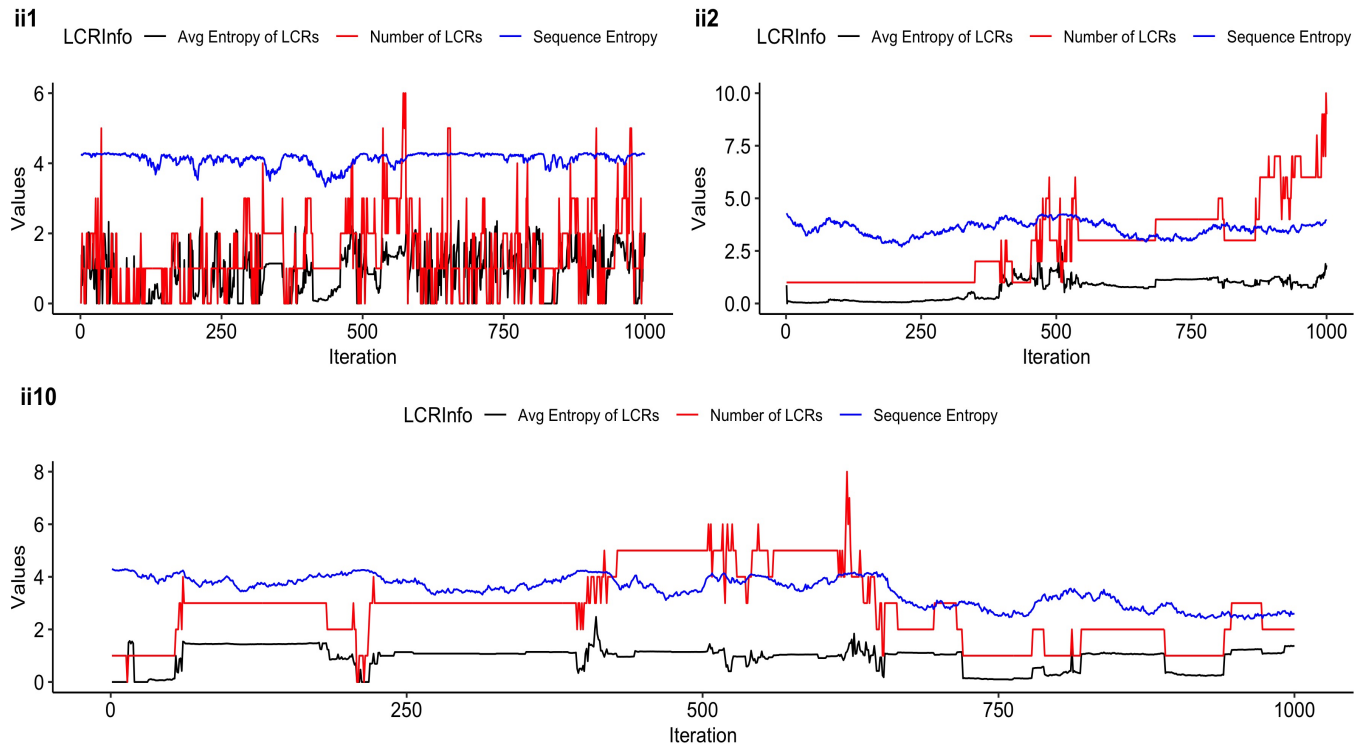


Figure 2: Entropy and number of LCR comparisons in a randomly generated protein sequence that has undergone simulated evolution. (TOP LEFT) An indel rate of 1 was utilized. (TOP RIGHT) An indel rate of 2 was utilized. (BOTTOM) An indel rate of 10 was utilized. A mutation rate of 1 was used for all.

To explore the effects of mutation on LCR formation, Figure 3 and Figure 4 show the average sequence entropy, the entropy of the LCRs in the sequence, and the total number of LCRs found in a randomly generated protein sequence that has undergone simulated mutation for 1000 generations. It is expected that under higher rates of mutation, indels will be difficult to form as the large number of mutations destroy repetitive segments. Figure 3 utilized mutation rates of 0.01, 0.1 and 0.5, all of which are less than the fixed indel rate of 1. In all three cases, the average sequence entropy fluctuates as the number of LCRs remains fairly large. In Figure 4, mutation rates are increased to 1, 2, and 10, much larger than the fixed indel rate. The average sequence entropy fluctuates much less in these cases and the number of LCRs in the sequence decreases in comparison to Figure 3. These results match the expectation that a larger mutation rate will have major effects of insertions and deletions. In the case where we set the mutation rate to be 10x the indel rate, only three LCRs are formed in total and are quickly removed due to mutation. In this case the entropy is also relatively unaffected, meaning that repetitive regions are not being formed.

It should also be noted that the average entropy of the low complexity regions being formed is significantly less than the average entropy of the entire sequence, and this is expected.

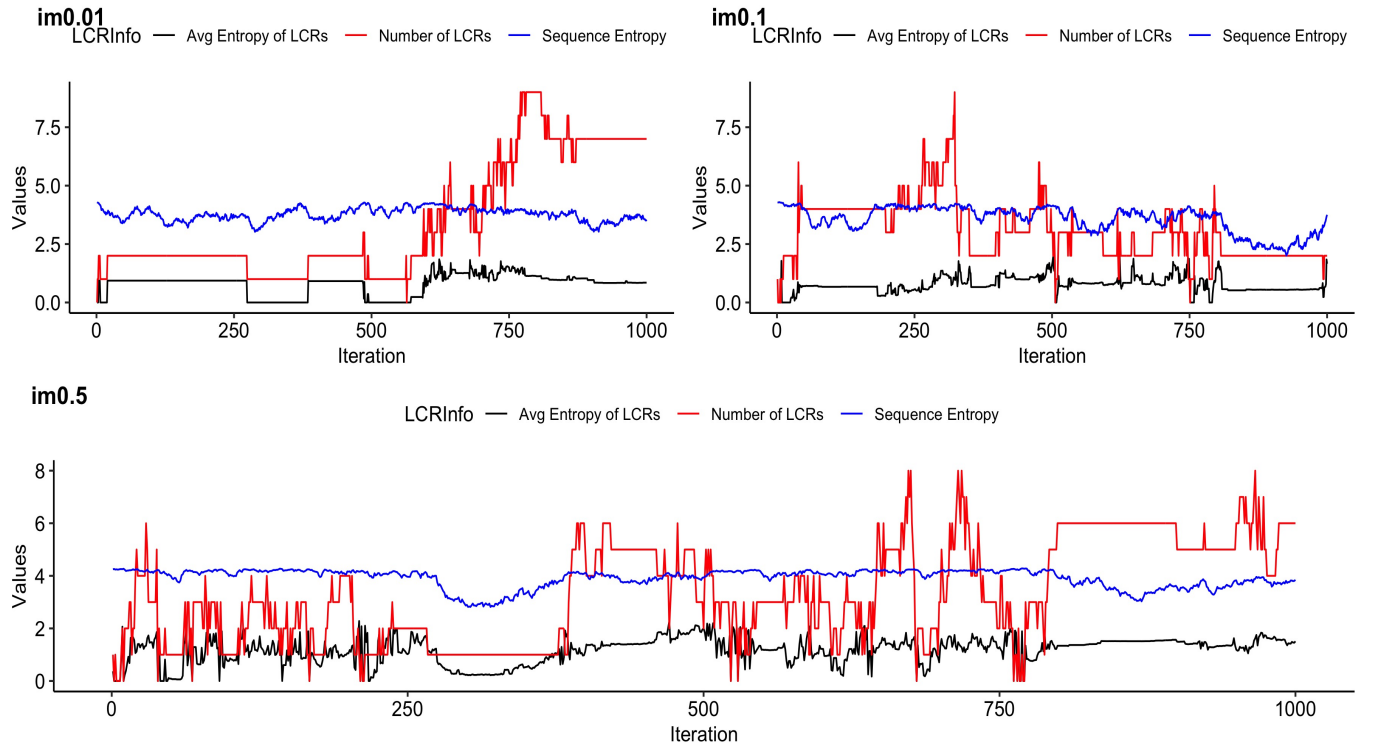


Figure 3: Entropy and number of LCR comparisons in a randomly generated protein sequence that has undergone simulated evolution. (TOP LEFT) A mutation rate of 0.01 was utilized. (TOP RIGHT) A mutation rate of 0.1 was utilized. (BOTTOM) A mutation rate of 0.5 was utilized. An indel rate of 1 was used for all.

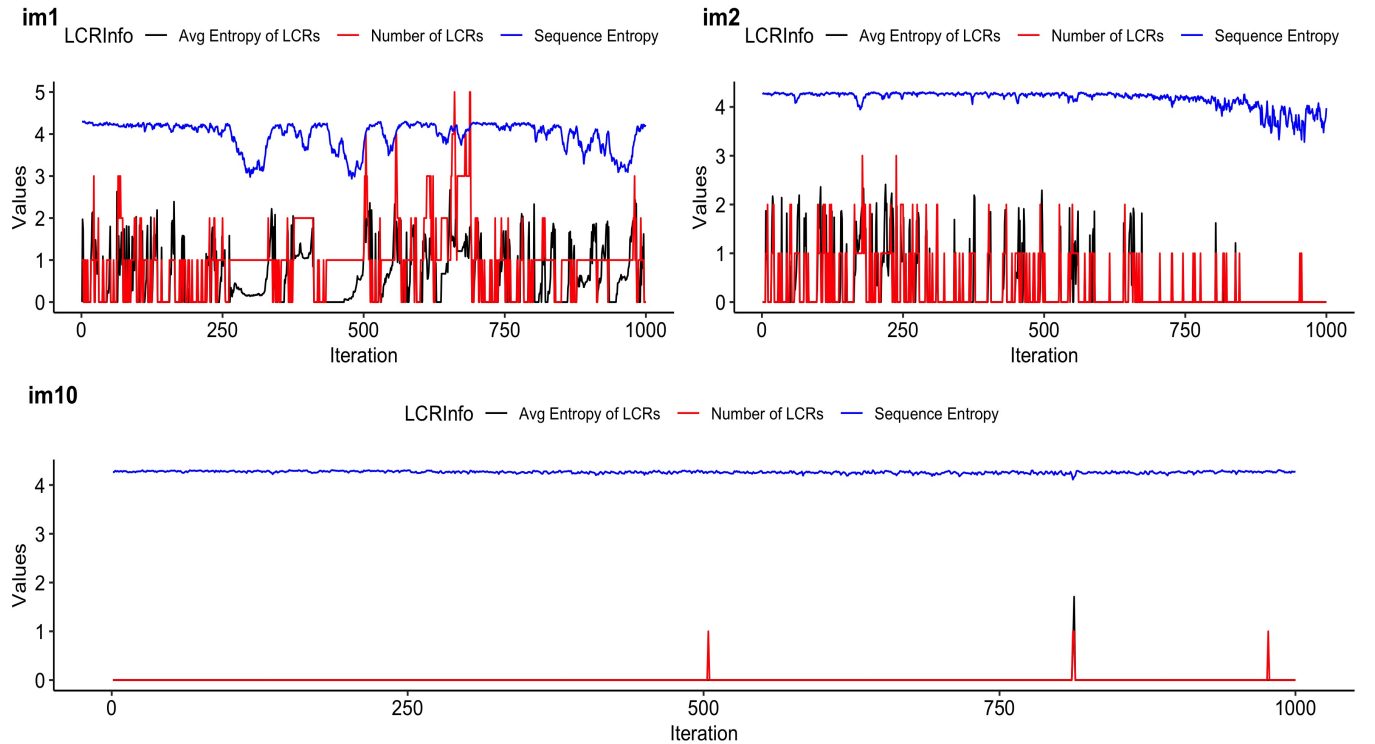


Figure 4: Entropy and number of LCR comparisons in a randomly generated protein sequence that has undergone simulated evolution. (TOP LEFT) A mutation rate of 1 was utilized. (TOP RIGHT) A mutation rate of 2 was utilized. (BOTTOM) A mutation rate of 0.5 was utilized. An indel rate of 1 was used for all.

Applying the Simulator to an ABC-MCMC

The second part of this study attempted to create an ABC-MCMC, using the LCR evolution simulator as a step in the algorithm. Time constraints in programming the ABC did not allow for the visualization of posterior distributions. We did however establish that the evolution simulator must be run for a minimum 1000 iterates for each newly proposed parameter value, in order to reach an almost "equilibrium" state. Figure 5 shows the euclidean distance plotted over each iteration of the simulation. The euclidean distance was calculated using vectors of summary statistics for both the simulated protein sequence and the observed protein sequence (SRP40 *S. cerevisiae*). Essentially the larger the distance, the more differences there are between the simulated protein sequence and the SRP40 protein sequence. In the plot on the left, mutation rate is set to 0 and indel rate is set to 1. In this case, we see the distance between the mutated protein and SRP40 increasing quite rapidly as more LCRs are being formed, and then start to reach a more steady state around 0.5. In the plot on the right, mutation and indel rates are both set to 1. In this case, since mutations are involved as well now, we see the distance rapidly increase and start to decrease more steadily at around 0.85.

For each newly proposed parameter value in the ABC-MCMC, we run through the evolution simulator 1000 times and take the average of all 1000 vectors produced. This is why we need to reach some sort of equilibrium in distance to ensure that we reach a level status of both parameters being proposed.

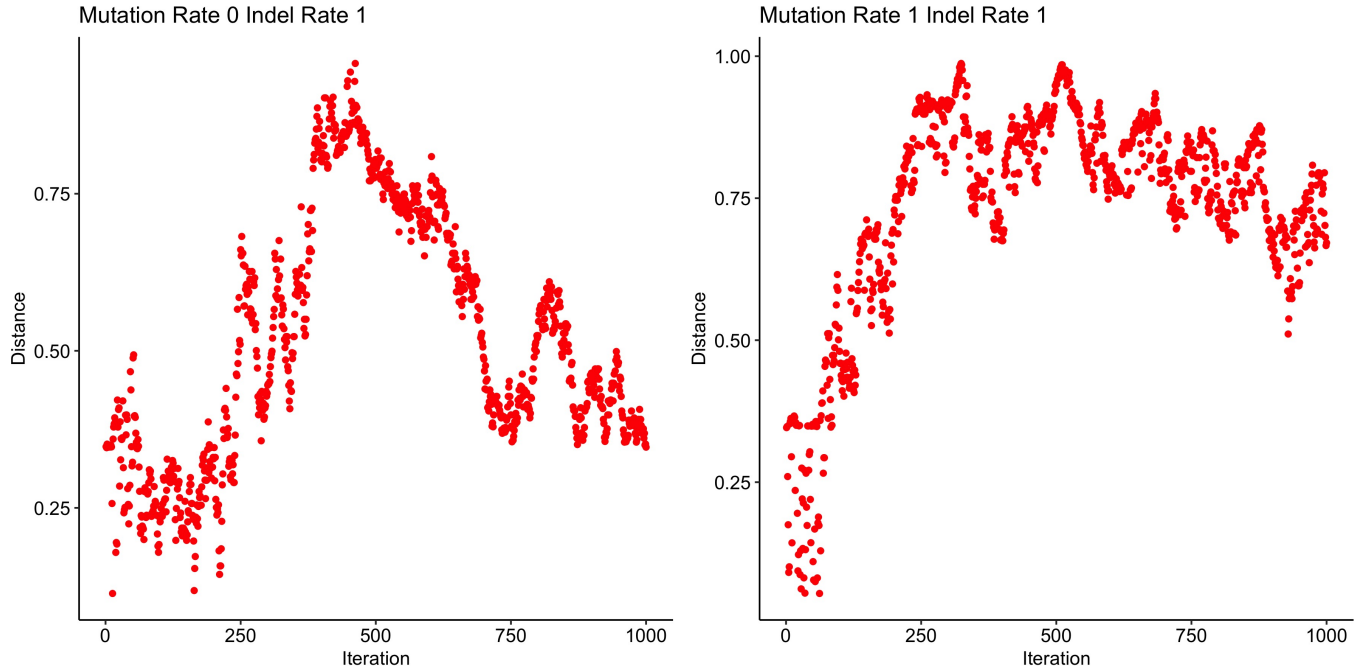


Figure 5: Calculated euclidean distances between mutated protein and SRP40 protein over 1000 evolution simulator iterates. (LEFT) mutation rate 0, indel rate 1. (RIGHT) mutation rate 1, indel rate 1.

Discussion

This study has developed a method of LCR simulation, which can be used in combination with an approximate bayesian computation markov chain monte carlo method. Due to timing, this study lacks estimates of parameter values that correspond to mutation and indel rates. However, this study does show the benefits of utilizing an approach that does not require calculation of the likelihood function. In the case of this study, mutations, insertions, and deletions alter the landscape of a protein sequence, making it extremely challenging to calculate the likelihood function. In the context of population genetics, ecology, epidemiology and systems biology, ABC approaches have been rapidly gaining popularity due to the massive increase in both the quantity and complexity of biological data (Sunnåker et al. 2013). With the advent of technology and data collection, we are at a point in time where new, tractable approaches must be developed and tested.

Mutations Have the Ability to Destroy LCR Formation

Results obtained from the evolution simulator are heavily dependent on the selected mutation and indel rates. Although both are forms of mutation, each rate parameter corresponds to a different process underneath the hood. Mutation rate in this case refers to the rate at which a base substitution occurs (changing one amino acid into another), while the indel rate refers to the rate at which an insertion or deletion takes place (adding or removing an amino acid residue). By altering mutation and indel rates, we are able to prove the simulation is functioning, and qualitatively assess the effects of having more point mutations on indel formation and vice versa.

With indel rates of 0.1 and 0.5 and a mutation rate of 1, very few low complexity regions are formed, and tend to be destroyed very easily. This is potentially due to the mutation rate being larger, as point mutations have the ability to destroy repetitive regions. As the indel rate is increased to 1, 2, and 10, while still holding the mutation rate constant at 1, LCR's are formed in much higher quantities and remain in the sequence for much longer. The overall sequence entropy also decreases with these rates more than the previous rates, due to the increase of low complexity regions contributing to the decrease in sequence entropy. It is interesting to observe that an indel rate 10x larger than the mutation rate does not seem to produce more repetitive, low complexity regions, that swamp out the mutations being created. This could be due in part to the fact that since deletions are also simultaneously occurring, this process could affect how many LCRs are created.

This is not the case however when mutation rates are adjusted and indel rates are held constant at 1. With mutation rates of 0.01, 0.1, and 0.5, many low complexity regions are being formed and the overall sequence entropy tends to fluctuate rapidly. Since the indel rate is higher than the mutation rate in this case, LCRs will be formed and remain stable throughout the entire simulation. When mutation rates are increased to values larger than the indel rates, very few LCRs will be formed and the average sequence entropy will remain close to maximum at about 4.3. With a mutation rate 10x larger than the indel rate, we see the formation of only 3 LCRs which are destroyed in the next iteration, and the average sequence entropy does not

fluctuate. This suggests that when the mutation rate is much larger than the indel rate, we are very rarely forming repeats, and when we do, they seem to be immediately destroyed by mutation.

Does the Length of a Repeat Play a Role in Insertions/Deletions

Insertions and deletions (indels) are a very important type of spontaneous mutation, as they make up the majority of the divergence among species (Britten 2002; Anzai et al. 2003). Indels however have been studied much less extensively compared to point mutations (Cartwright 2009). In this study, deviates were drawn from the exponential distribution and assigned to each individual amino acid residue. When deviates were assigned to residues that were part of a repetitive unit, the length of the repeat was multiplied by the indel rate, and this value was used as the scale parameter (β) of the exponential distribution. The reason for doing this is because it has been repeatedly observed in both DNA and proteins that the occurrence of indels declines monotonically as a function of their length (Pascarella and Argos 1992; Benner et al. 1993; Qian and Goldstein 2001; Loewenthal et al. 2021). As the length of a repeat increases, so does the scale parameter we use to draw deviates from the exponential distribution. With a larger scale parameter, the mean of the distribution also increases and there is a greater chance we draw deviates with larger values. This subsequently means that the longer the repeat is, the less likely it is to continue to gain repeats and hence we see this monotonic decrease in the occurrence of indels.

This monotonic decrease could also explain why an indel rate 10x larger than the mutation rate does not produce a protein chalked full of repetitive regions. As the indel rate is increased, the number of LCRs increases, but as these LCRs continue to increase in length, it becomes increasingly rare to continue to add repeats in that region. An ideal extension of this research would be to explore alternative relationships which describe the role of the length of a repetitive region in its subsequent mutation. This could be done through assigning deviates for repeats based on different mathematical functions. Currently deviates are assigned to repeated amino acids by multiplying the length of the repeat by the indel rate, but this could be potentially changed to multiplying the indel rate by an exponential function, e^{length} . In this case, as the length of the repetitive region increases, the exponential part of the function would also increase exponentially. This means that when we have longer repeats, it will be more difficult to keep adding repeats due to the extremely large scale parameter that would be used to draw deviates from the exponential distribution.

Why Parameters Could not be Estimated

As previously stated, the ABC-MCMC did not produce posterior distributions to estimate parameter values for mutation and indel rates. Overall, the process behind an ABC-MCMC is very complex, but in this study we simplified many of the steps in the algorithm, which could potentially be the reason posterior distributions were not produced.

An area of major concern in this study stemmed from the chosen summary statistics. Summary statistics are vital in

order to quantitatively capture relevant information about the data under analysis (Sunnåker et al. 2013). One potential issue with the summary statistics is that they could not be explaining the characteristics we want in the best way. In this study, we utilized the average number of LCRs, the average entropy of those LCRs, and the length of the entire protein sequence as characteristics that explained the protein sequence. These summary statistics may be too general to accurately capture the differences between a protein of known low complexity (SRP40 *S. cerevisiae*), and a randomly simulated protein sequence. The goal of the ABC-MCMC was to essentially simulate a protein similar to a known protein sequence using various mutation and indel rates, in order to estimate the rates that produce the known sequence. The way this simulated sequence is compared to the known sequence is only through these summary statistics, not by the composition of the sequence itself. This means for example that a simulated protein sequence could have 3 LCRs (Same as in SRP40), but these LCRs could be in different locations and made up of different amino acids, which means the simulated protein is still very different from the known protein, but we are inaccurately saying they are similar based on this summary statistic.

It could also be the case however that the summary statistics used in this study required specific weightings to minimize certain statistics from dominating the distance calculation. The length summary statistic tended to be a dominating force when running the ABC, which affected how often we accepted and rejected parameter values. Due to the fact that we attempted to estimate two parameters with the ABC, it is often the case that while a statistic may estimate one of these parameters well, it may not do the same for the other (Hamilton et al. 2005). A potential workaround for this is to implement a weighting scheme that essentially gives greater weights to statistics carrying more information on the parameter of interest (Hamilton et al. 2005). The way this could be done is to assess relationships between parameters and statistics using local regressions, and subsequently implement a weighted Euclidean distance using the results from the local regression. (Hamilton et al. 2005). Through multiplying summary statistics by a certain weight, it could potentially allow for the acceptance of more parameters due to statistics like length not being the dominating factor in the Euclidean distance calculation.

Conclusion

In summary, the creation of a program that simulates the evolution of low complexity regions can enable the exploration of new model based approaches such as approximate bayesian computations. This can subsequently allow for predictions to be made regarding the evolutionary history of protein or DNA sequences. This research also contributes to the lack of knowledge there is surrounding indels and the way in which they evolve. Through adjusting models of indel evolution using the simulator, the way in which indels are formed can be better understood. This research provides the ground work that is required for utilizing approximate bayesian computation approaches in the context of low complexity regions. The evolution of low complexity regions is still a growing area of research, as the complexity behind how they evolve makes it very difficult to analyze the way they evolve. This research can help pave the way for future work into using model based analyses for exploring the evolutionary dynamics of low complexity regions in proteins.

References

- Anzai T, Shiina T, Kimura N, Yanagiya K, Kohara S, Shigenari A, Yamagata T, Kulski J K, Naruse T K, Fujimori Y, et al. (2003). Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. *Proceedings of the National Academy of Sciences* 100(13), 7708–7713.
- Bannen R M, Bingman C A, and Phillips G N (2007). Effect of low-complexity regions on protein structure determination. *Journal of Structural and Functional Genomics* 8(4), 217–226.
- Battistuzzi F U, Schneider K A, Spencer M K, Fisher D, Chaudhry S, and Escalante A A (2016). Profiles of low complexity regions in Apicomplexa. *BMC evolutionary biology* 16(1), 1–12.
- Bayliss C D, Field D, Moxon E R, et al. (2001). The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. *The Journal of clinical investigation* 107(6), 657–666.
- Benner S A, Cohen M A, and Gonnet G H (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *Journal of molecular biology* 229(4), 1065–1082.
- Brinkmann B, Klintschar M, Neuhuber F, Hühne J, and Rolf B (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics* 62(6), 1408–1415.
- Britten R J (2002). Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences* 99(21), 13633–13635.
- Cartwright R A (2009). Problems and solutions for estimating indel rates and length distributions. *Molecular biology and evolution* 26(2), 473–480.
- Coletta A, Pinney J W, Solís D Y W, Marsh J, Pettifer S R, and Attwood T K (2010). Low-complexity regions within protein sequences have position-dependent roles. *BMC systems biology* 4(1), 1–13.
- DePristo M A, Zilversmit M M, and Hartl D L (2006). On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene* 378, 19–30.
- Dunker A K, Brown C J, Lawson J D, Iakoucheva L M, and Obradović Z (2002). Intrinsic disorder and protein function. *Biochemistry* 41(21), 6573–6582.
- Ellegren H (2004). Microsatellites: simple sequences with complex evolution. *Nature reviews genetics* 5(6), 435–445.
- Everett C and Wood N (2004). Trinucleotide repeats and neurodegenerative disease. *Brain* 127(11), 2385–2405.
- Gardner M J, Hall N, Fung E, White O, Berriman M, Hyman R W, Carlton J M, Pain A, Nelson K E, Bowman S, et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906), 498–511.
- Haerty W and Golding G B (2010a). Genome-wide evidence for selection acting on single amino acid repeats. *Genome research* 20(6), 755–760.
- Haerty W and Golding G B (2010b). Low-complexity sequences and single amino acid repeats: not just “junk” peptide sequences. *Genome* 53(10), 753–762.

- Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, and Excoffier L (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* 170(1), 409–417.
- Huntley M and Golding G B (2000). Evolution of simple sequence in proteins. *Journal of molecular evolution* 51(2), 131–140.
- Huntley M A and Clark A G (2007). Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Molecular biology and evolution* 24(12), 2598–2609.
- Huntley M A and Golding G B (2002). Simple sequences are rare in the Protein Data Bank. *Proteins: Structure, Function, and Bioinformatics* 48(1), 134–140.
- Huntley M A and Golding G B (2006). Selection and slippage creating serine homopolymers. *Molecular biology and evolution* 23(11), 2017–2025.
- Karlin S, Brocchieri L, Bergman A, Mrázek J, and Gentles A J (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proceedings of the National Academy of Sciences* 99(1), 333–338.
- Kimura M (1983). *The neutral theory of molecular evolution*. Cambridge University Press.
- Kruglyak S, Durrett R, Schug M D, and Aquadro C F (2000). Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Molecular Biology and Evolution* 17(8), 1210–1219.
- Kruglyak S, Durrett R T, Schug M D, and Aquadro C F (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences* 95(18), 10774–10778.
- Kumari B, Kumar R, and Kumar M (2015). Low complexity and disordered regions of proteins have different structural and amino acid preferences. *Molecular BioSystems* 11(2), 585–594.
- Laurin-Lemay S, Dickson K, and Rodrigue N (2022). Jump-Chain Simulation of Markov Substitution Processes Over Phylogenies. *Journal of Molecular Evolution*, 1–5.
- Levinson G and Gutman G A (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular biology and evolution* 4(3), 203–221.
- Liepe J, Barnes C, Cule E, Erguler K, Kirk P, Toni T, and Stumpf M P (2010). ABC-SysBio—approximate Bayesian computation in Python with GPU support. *Bioinformatics* 26(14), 1797–1799.
- Loewenthal G, Rapoport D, Avram O, Moshe A, Wygoda E, Itzkovitch A, Israeli O, Azouri D, Cartwright R A, Mayrose I, et al. (2021). A probabilistic model for indel evolution: differentiating insertions from deletions. *Molecular biology and evolution* 38(12), 5769–5781.
- Marcotte E M, Pellegrini M, Yeates T O, and Eisenberg D (1999). A census of protein repeats. *Journal of molecular biology* 293(1), 151–160.
- Marjoram P (2013). Approximation bayesian computation. *OA genetics* 1(3), 853.
- Marjoram P, Molitor J, Plagnol V, and Tavaré S (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 100(26), 15324–15328.

- Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, Gruca A, Plewczynski D, Grynberg M, Bernadó P, et al. (2020). Disentangling the complexity of low complexity proteins. *Briefings in Bioinformatics* 21(2), 458–472.
- Mirkin S M (2007). Expandable DNA repeats and human disease. *Nature* 447(7147), 932–940.
- Mularoni L, Veitia R A, and Albà M M (2007). Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* 89(3), 316–325.
- Nevo E (2001). Genetic diversity.
- Newfeld S J, Smoller D A, and Yedvobnick B (1991). Interspecific comparison of the unusually repetitive *Drosophila* locus-mastermind. *Journal of molecular evolution* 32(5), 415–420.
- Newfeld S J, Tachida H, and Yedvobnick B (1994). Drive-selection equilibrium: homopolymer evolution in the *Drosophila* gene mastermind. *Journal of molecular evolution* 38(6), 637–641.
- Ohta T and Kimura M (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetics Research* 22(2), 201–204.
- Pascarella S and Argos P (1992). Analysis of insertions/deletions in protein structures. *Journal of molecular biology* 224(2), 461–471.
- Qian B and Goldstein R A (2001). Distribution of indel lengths. *Proteins: Structure, Function, and Bioinformatics* 45(1), 102–104.
- Ross C A, McInnis M G, Margolis R L, and Li S.-H (1993). Genes with triplet repeats: candidate mediators of neuropsychiatric disorders. *Trends in neurosciences* 16(7), 254–260.
- Shannon C E (1948). A mathematical theory of communication. *The Bell system technical journal* 27(3), 379–423.
- Sunnåker M, Busetto A G, Numminen E, Corander J, Foll M, and Dessimoz C (2013). Approximate bayesian computation. *PLoS computational biology* 9(1), e1002803.
- Verstrepen K J, Jansen A, Lewitter F, and Fink G R (2005). Intragenic tandem repeats generate functional variability. *Nature genetics* 37(9), 986–990.
- Warren S T, Muragaki Y, Mundlos S, Upton J, and Olsen B R (1997). Polyalanine expansion in synpolydactyly might result from unequal crossing-over of *HOXD13*. *Science* 275(5298), 408–409.
- Wootton J C and Federhen S (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry* 17(2), 149–163.
- Wright P E and Dyson H J (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology* 16(1), 18–29.
- Wu S H and Rodrigo A G (2015). Estimation of evolutionary parameters using short, random and partial sequences from mixed samples of anonymous individuals. *BMC bioinformatics* 16(1), 1–12.

The following files included these appendices are located at /home/alex/scratch/abcmcmc-thesis4cl2

Appendix 1: Low Complexity Region Evolution Simulator

/home/alex/scratch/abcmcmc-thesis4cl2/mutations_2_BG_vecs.cpp

```

1 #include "functions.cpp"
2 #include "getindex.cpp"
3 // #include "simulated_protein.cpp"
4 #include <bits/stdc++.h>
5 #include <iostream>
6 #include <vector>
7 #include <string>
8 #include <random>
9 #include <ctime>
10 #include <cmath>
11 #include <map>
12 #include <algorithm> //this is to get min element stuff, cool library
13 #define numAA 20
14 using namespace std;
15
16 Ran myran(time(NULL)); //We will use 21 as the random seed right now, used in Poissondev too
17
18 ///////////////////////////////////////////////////
19 // FIRST FUNCTION TAKES AN INTEGER VALUE AND GENERATES A RANDOM/
20 // AMINO ACID SEQUENCE OF THAT LENGTH /
21 ///////////////////////////////////////////////////
22 std::string createSeq(int n){
23
24     char aminoAcids[numAA] = { 'G', 'A', 'L', 'M', 'F', 'W', 'K', 'Q', 'E', 'S', 'P', 'V',
25                                'I', 'C', 'Y', 'H', 'R', 'N', 'D', 'T' };
26
27     std::string protein = "";
28     for (int i = 0; i < n; i++){
29         protein += aminoAcids[myran.int64() % numAA];} //this rand() % 20 means in the range 0-19
30
31     //std::cout << protein << "\n" << "\n" ;
32     return protein;
33 }
34
35 ///////////////////////////////////////////////////
36 // FUNCTION TO GENERATE NORMAL DEVIATES //
37 ///////////////////////////////////////////////////

```



```

38 double getNormalDev(double mu, double stdev) {
39     Normaldev mynorm(mu, stdev, myran.int64());
40     double dev = mynorm.dev();
41     //std::cout << dev << "\n";
42     return dev;
43 }
44
45 //Trying different function to generate random deviates
46 double getNormalDev2(double mu, double stdev) {
47     std::random_device rd;
48     std::mt19937 gen(rd());
49     std::normal_distribution<double> dist(mu, stdev);
50
51     double random_num = dist(gen);
52     return random_num;
53 }
54
55 //using uniform real distribution to generate numbers between
56 //0.0005 and 0.5 for the ttest pvalue
57
58 double random_num_zero_half() {
59     std::random_device rd;
60     std::mt19937 gen(rd());
61     std::uniform_real_distribution<double> dist(0.0,0.5);
62
63     double ran_num = dist(gen);
64     return ran_num;
65 }
66
67 //FUNCTION TO CALCULATE ENTROPY OF PROTEIN SEQUENCE
68 double calc_entropy(string protein_seq) {
69     map<char, int> freq_map;
70     for (char c : protein_seq) {
71         freq_map[c]++;
72     }
73     double entropy = 0.0;
74     int seq_len = protein_seq.length();
75     for (auto const& pair : freq_map) {
76         double prob = (double) pair.second / seq_len;
77         entropy -= prob * log2(prob);
78     }
79     return entropy;
80 }

```

```

81
82 //
    ///////////////////////////////////////////////////////////////////

83 /////////////// Changed here down - Brian - Feb 12 Xander made Changes, trying to understand the new stuff
    ///////////////

84 //
    ///////////////////////////////////////////////////////////////////

85 std::string mutateSeqExpBG(std::string simulated_protein, double mutation_rate, double indel_rate){ //
    took out iterates as a parameter

86
87 //Setting up the vectors
88 std::vector<double> mut_dev;
89 std::vector<double> ind_dev;
90
91 char aminoAcids[20] = { 'G', 'A', 'L', 'M', 'F', 'W', 'K', 'Q', 'E',
92     'S', 'P', 'V', 'I', 'C', 'Y', 'H', 'R', 'N', 'D', 'T' };
93
94 //std::cout << "before mutateseqEXP:\t" << simulated_protein << "\n"; // Initially printing the non-
    mutated strin.
95 int len = simulated_protein.length();
96
97 //ASSIGN DEVIATES FOR MUTATION
98 for (int i = 0; i < len; i++) {
99     float beta1 = mutation_rate ;
100     Expondev myexp(beta1,myran.int64());
101     double deviate = myexp.dev(); // choose exp_deviate(mean of beta)
102     mut_dev.push_back(deviate);
103 }
104 // Traverse the string and generate deviates FOR INDELS - this seems to be
105 // working, counter is showing correct values for repeats -Feb 15
106 for (int i = 0; i < len; i++) {
107     int counter = 1 ;
108     //Code to scan back and forth to find repeats - this i-1 will give -1 on the first iteration
    though - check this in while loop too for scanning backwards
109     if (simulated_protein[i] != simulated_protein[i+1] && simulated_protein[i] != simulated_protein[i
    -1]) {
110         float beta2 = indel_rate ; // the length if no repeats is 1
111         Expondev myexp(beta2,myran.int64());
112         double deviate = myexp.dev();
113         ind_dev.push_back(deviate);
114     } else {

```

```

115     int x = 1 ;
116     int y = 1 ;
117     //Be careful in these while loops, for i-y, when i is 0
118     //and y is 1, how does it not throw error
119     //Looking forward for repeats
120     while (simulated_protein[i] == simulated_protein[i + x]) {
121         counter += 1 ;
122         x++;
123     }
124     //Looking backwards for repeats
125     while (simulated_protein[i] == simulated_protein[i - y]) {
126         counter += 1 ;
127         y++;
128     }
129     float beta3 = indel_rate * counter ; // trying to see if the # of repeats plays a role,
multiply by counter
130     Expondev myexp(beta3,myran.int64());
131     double deviate = myexp.dev();
132     ind_dev.push_back(deviate);
133 }
134 //std::cout << counter << "\n"; //Checking if counter is finding repeats
135 }
136
137 // Working with the vector of structs, why we doing 0.5*len
138 // FEB 15 - WORKING WITH VECTORS NOW
139 for(int iter=0; iter<0.5*len; iter++) { // throw down 0.5 mut/site
140     //selecting the lowest deviate from both vectors
141     double minExpDev=mut_dev[0];
142     int minPosition=0;
143     int minType=0; // type=0 for mut, 1 for ins, 2 for del
144     for (int i = 0; i < len; i++) {
145         if(minExpDev > mut_dev[i]) {
146             minExpDev=mut_dev[i];
147             minPosition=i; minType=0;
148         }
149         if(minExpDev > ind_dev[i]) {
150             minExpDev=ind_dev[i];
151             minPosition=i; minType=1;
152             if(myran.doub() < 0.5) { minType=2; }
153         }
154     }
155     //std::cout << "minpos" << minPosition << "\n";
156     //std::cout << "mindev" << minExpDev << "\n";

```

```

157     //std::cout << "mintype" << minType << "\n";
158     //std::cout << "iter" << iter << "\n";
159
160     //This is where base changes, insertions, deletions occur
161     if(minType==0) { // put in a base change
162         simulated_protein[minPosition]=aminoAcids[myran.int64() % numAA];
163         float betal = mutation_rate ;
164         Expondev myexp(betal,myran.int64());
165         double deviate = myexp.dev();
166         mut_dev[minPosition]=deviate;//swap deviate for base change
167         //Printing to see mutated protein afer base change//
168         //std::cout << "\n" << "after basechange:\t" << simulated_protein << "\n" << "\n" ;
169         // THIS IS JUST TO PRINT THE VECTOR OF DEVIATES FOR MUTATION TO SEE IF IT CHANGED
170         /*for (int x = 0; x < mut_dev.size(); x++) {
171             std::cout << mut_dev[x] << ' ';
172         }
173         for (int x = 0; x < ind_dev.size(); x++) {
174             std::cout << ind_dev[x] << ' ';
175         }*/
176     } else { // put in an indel
177         if(minType==1) { // use insertion
178             if(len < 1.5*len) {
179                 len++;
180                 simulated_protein.insert(minPosition+1, 1, simulated_protein[minPosition]); // insert
repeat beside minPosition
181                 float betal = mutation_rate ;
182                 Expondev myexp(betal,myran.int64());
183                 double deviate = myexp.dev();
184                 mut_dev.insert(mut_dev.begin() + minPosition+1, deviate); //Adding a new deviate to
mutation vector
185                 ind_dev.insert(ind_dev.begin() + minPosition+1, deviate); //Adding new deviate to
indel vector as placeholder
186                 //Printing to see mutated protein after insertion//
187                 //std::cout << "\n" << "after insertion:\t" << simulated_protein << "\n" << "\n" ;
188                 // THIS IS JUST TO PRINT THE VECTOR OF DEVIATES FOR MUTATION TO SEE IF IT CHANGED
189                 /*for (int x = 0; x < mut_dev.size(); x++) {
190                     std::cout << mut_dev[x] << ' ';
191                 }
192                 for (int x = 0; x < ind_dev.size(); x++) {
193                     std::cout << ind_dev[x] << ' ';
194                 }*/
195             }
196         } else { // use deletion

```

```

197         if(len > 0.5*len) {
198             len--;
199             simulated_protein.erase(minPosition, 1); //delete amino acid at minPosition
200             mut_dev.erase(mut_dev.begin() + minPosition); //Deleting deviate in the mutation
vector
201             ind_dev.erase(ind_dev.begin() + minPosition); //Deleting deviate in the indel vector
202             //Printing to see mutated protein after deletion//
203             //std::cout << "\n" << "after deletion:\t" << simulated_protein << "\n" << "\n" ;
204             // THIS IS JUST TO PRINT THE VECTOR OF DEVIATES FOR MUTATION TO SEE IF IT CHANGED
205             /*for (int x = 0; x < mut_dev.size(); x++) {
206                 std::cout << mut_dev[x] << ' ';
207             }
208             for (int x = 0; x < ind_dev.size(); x++) {
209                 std::cout << ind_dev[x] << ' ';
210             }*/
211         }
212     }
213 }
214
215 // THIS CHUNK OF CODE CAUSING MEMORY ISSUES, INDEXING SHIT - FEB 17
216 // check what is around the change
217 // need to look down one, the site and up one
218 // So here were checking if the insertion,
219 // or deletion affected the landscape of the DNA sequence.
220 // Did it create a repeat, interrupt a repeat, etc.
221
222 //MUTATION AT START OF SEQUENCE
223 if (minPosition == 0) {
224     if (simulated_protein[minPosition] == simulated_protein[minPosition+1]) {
225         int counter = 1;
226         int x = 1;
227         while (simulated_protein[minPosition] == simulated_protein[minPosition+x]) {
228             counter += 1;
229             x++;
230         }
231         int start = 0;
232         int end = x-1;
233         //std::cout << "start" << start << "\n";
234         //std::cout << "end" << end << "\n";
235         float beta3 = indel_rate * counter ;
236         for(int j=start; j<=end; j++) {
237             Expondev myexp(beta3,myran.int64());
238             double deviate = myexp.dev();

```

```

239         ind_dev[j]=deviate;
240     }
241 } else { //NOT PART OF REPEAT
242     float beta3 = indel_rate;
243     Expondev myexp(beta3,myran.int64());
244     double deviate = myexp.dev();
245     ind_dev[minPosition]=deviate;
246 }
247 } else if (minPosition == simulated_protein.length()) { //Mutation at end of sequence
248     if (simulated_protein[minPosition] == simulated_protein[minPosition-1]){
249         int counter = 1;
250         int y = 1;
251         while (simulated_protein[minPosition] == simulated_protein[minPosition-y]) {
252             counter += 1;
253             y++;
254         }
255         int start = minPosition - y;
256         int end = simulated_protein.length();
257         //std::cout << "start" << start << "\n";
258         //std::cout << "end" << end << "\n";
259         float beta3 = indel_rate * counter;
260         for(int j=start; j<=end; j++) {
261             Expondev myexp(beta3, myran.int64());
262             double deviate = myexp.dev();
263             ind_dev[j]=deviate;
264         }
265     } else { //NOT PART OF REPEAT
266         float beta3 = indel_rate;
267         Expondev myexp(beta3, myran.int64());
268         double deviate = myexp.dev();
269         ind_dev[minPosition]=deviate;
270     }
271 } else { //mutation in middle of sequence
272     if (simulated_protein[minPosition] == simulated_protein[minPosition+1] || simulated_protein[
minPosition] == simulated_protein[minPosition-1]) { // part of a repeat
273         int counter = 1;
274         int x = 1;
275         int y = 1;
276         //Look backward
277         while (simulated_protein[minPosition] == simulated_protein[minPosition-y]) {
278             counter += 1;
279             y++;
280         }

```

```

281     int start = minPosition - y; //Getting start position of repeats
282     //std::cout << "start" << start+1 << "\n";
283     //count forward
284     while (simulated_protein[minPosition] == simulated_protein[minPosition + x]) {
285         counter += 1;
286         x++;
287     }
288     int end = minPosition + x; //Getting end position of repeats
289     //std::cout << "end" << end-1 << "\n";
290     float beta3 = indel_rate * counter ;
291     for(int j=start+1; j<=end-1; j++) {
292         Expondev myexp(beta3,myran.int64());
293         double deviate = myexp.dev();
294         ind_dev[j]=deviate;
295     }
296 } else { // not part of a repeat
297     float beta3 = indel_rate;
298     Expondev myexp(beta3,myran.int64());
299     double deviate = myexp.dev();
300     ind_dev[minPosition]=deviate;
301 }
302 }
303 }
304 //std::cout << "\n" << "after mutateSeqEXP:\t" << simulated_protein << "\n" << "\n" ;
305 return simulated_protein;
306 }
307
308 ///// TRYING TO DEBUG THIS FILE WITH THIS MAIN FUNCTION /////
309 /*int main() {
310     std::string simulated_protein = createSeq(10);
311     for (int j = 0; j < 5; j++){
312         // std::string simulated_protein = createSeq(10);
313         double mutation_rate = 0.14;
314         double indel_rate = 0.14;
315         std::cout << calc_entropy(simulated_protein) << "\n";
316         std::string mutated_protein = mutateSeqExpBG(simulated_protein, mutation_rate, indel_rate);
317         simulated_protein = mutated_protein;
318         sim_protein(simulated_protein);
319     }
320 }*/

```

Appendix 2: ABC-MCMC in C++

/home/alex/scratch/abcmcmc-thesis4c12/abcmcmc.cpp

```

1 #include "mutations_2_BG_vecs.cpp"
2 #include "observed_protein.cpp"
3 #include "simulated_protein.cpp"
4 #include "distance2.cpp"
5 #include "vecAvg.cpp"
6 #include "ttest.cpp"
7 #include <stdio.h>
8 #include <stdlib.h>
9 #include <string.h>
10 #include <iostream>
11 #include <bits/stdc++.h>
12 #include <fstream> //The fstream library allows us to work with files
13 #include <vector> // Working with vectors is nicer than arrays, more c++
14
15 ///////////////////////////////////////////////////
16 // THE MAIN FUNCTION - WHERE EXECUTION BEGINS/
17 ///////////////////////////////////////////////////
18
19 int main() {
20
21     // Setting initial parameters
22     double mutation_rate = 10;
23     double indel_rate = 1;
24     int num_simulations = 1;
25     int num_mutations = 1;
26     double mean_proposal = 0.0;
27     double stddev_proposal = 1.0;
28
29     double mut_rate_arr[10000]; //for accepted mutation rates
30     double ind_rate_arr[10000]; //for accepted indel rates
31     double index[10000]; // simulation iterations
32     double distance_array[10000]; //for accepted distances
33
34     double current_mut_rate_arr[10000]; //for current mutation rates
35     double proposed_mut_rate_arr[10000]; //for proposed mutation rates
36     double current_ind_rate_arr[10000]; //for current indel rates
37     double proposed_ind_rate_arr[10000]; //for proposed indel rates
38     double current_distances_arr[10000]; //for current distances avg of 10
39     double proposed_distances_arr[10000]; //for proposed distances avg of 10

```



```

40  double distances_ttest[10000]; //for each of the 10 distances for proposed parameters
41
42  double acc_rej_rate[10000]; //acceptance-rejection rate for mutation
43
44  double mean_ttest_arr[10000]; //storing means from ttest of 10 vectors
45  double stddev_ttest_arr[10000]; //storing stdev from 10 vectors
46  double tstat_array[10000]; //storing test statistics
47  double probability_ttest[10000]; //storing pvalues for ttest
48  double accepted_rejected[10000]; //1 for accepted, 0 for rejected
49  double entropy_vec[10000];
50  std::vector<vector<double>> vec_of_vecs3; //for getting vectors of 10 distances
51  std::vector<vector<double>> vec_of_vecs4; //summary statistics for each of 10
52
53  std::vector<vector<double>> vec_of_vecs_length;
54  std::vector<vector<double>> vec_of_vecs_avgnum;
55  std::vector<vector<double>> vec_of_vecs_entropy;
56  std::vector<double> distance_current_vec;
57
58  double new_mut_rate = 0;
59  double new_ind_rate = 0;
60  double acceptance_counter = 0;
61
62  //importing the vector of summary statistic for observed protein SRP40 (Saccharomyces)
63  //AND NORMALIZING IT, IT DOES NOT CHANGE
64  std::vector<double> obs_prot_vtr = og_protein();
65  obs_prot_vtr = normalize_vector(obs_prot_vtr);
66
67  //Generating current state protein sequence outside loop for
68  //first time
69  std::string simulated_protein = createSeq(1000); //For the current state
70
71  // GETTING THE FIRST CURRENT DISTANCE OUTSIDE THE MAIN LOOP
72  // TRYING TO GET MUTATED SEQUENCE HERE SO WE CAN KEEP MUTATING
73  // THE SAME SEQUENCE IN THE ALGORITHM
74
75  for (int a = 0; a < 1000; a++) {
76      for (int j = 0; j < num_mutations; j++){
77          std::string mutated_protein = mutateSeqExpBG(simulated_protein, mutation_rate, indel_rate);
78          simulated_protein = mutated_protein;
79      }
80
81      std::vector<double> sim_prot_vtr = sim_protein(simulated_protein);
82      sim_prot_vtr = normalize_vector(sim_prot_vtr);

```

```

83     double distance_current = vectors_distance2(sim_prot_vtr, obs_prot_vtr);
84     distance_current_vec.push_back(distance_current);
85 }
86
87 double distance_current = Mean(distance_current_vec);
88
89 std::cout << distance_current << '\n';
90
91 //std::string simulated_protein2 = createSeq(400); //for proposed state - going to use the same
sequence and keep mutating it for all 10 iterates below - keep mutating same sequence the whole
simulation
92
93 //First for loop is for the number of simulations
94 for (int i = 0; i < num_simulations; i++) {
95
96     //setting this in the loop
97     std::vector<vector<double>> vec_of_vecs; //for the current state
98     std::vector<vector<double>> vec_of_vecs2; //for proposed state
99     std::vector<double> distances; //holding vector of 10 distances per simulation iteration
100    std::vector<double> length_vec; //holding length summary statistic
101    std::vector<double> length_vec2; //holding unnormalized length
102    std::vector<double> num_lcrs_vec; //holding number of LCRs summary statistic
103    std::vector<double> avg_entropy_vec; //holding average entropy summary statistic
104
105    current_mut_rate_arr[i] = mutation_rate; // adding current mut rate to array
106    current_ind_rate_arr[i] = indel_rate; // adding current ind rate to array
107
108    // 0. Proposing new parameter values - this if statement is
109    // to hold one of the parameters constant on every other
110    // iteration... in other words, holding one parameter
111    // constant and changing the other on every other
112    // iteration. - I changed this back to just generating new
113    // ones each time.
114    new_mut_rate = getNormalDev2(mean_proposal, stddev_proposal) + mutation_rate; //+ mutation_rate;
115    std::cout << "Mutation rate: " << new_mut_rate << "\n";
116    new_ind_rate = getNormalDev2(mean_proposal, stddev_proposal) + indel_rate; //+ indel_rate;
117    std::cout << "Indel Rate: " << new_ind_rate << "\n";
118
119    if (i == 0) {
120        new_mut_rate = mutation_rate;
121        new_ind_rate = indel_rate;
122    }
123

```

```

124     proposed_mut_rate_arr[i] = new_mut_rate; // adding the proposed mut rate to array
125     proposed_ind_rate_arr[i] = new_ind_rate; // adding the proposed ind rate to array
126
127     std::string simulated_protein2 = createSeq(400); //for proposed state - going to use the same
sequence and keep mutating it for all 10 iterates below
128
129     // 1. We need to generate the random protein sequence
130     // For loop here is to generate 10 vectors of summary
131     // statistics for each parameter and get the average of all
132     for (int k = 0; k<1000; k++){
133         // 2. Next we need to mutate the simulated protein
134         // Going to try and mutate over 2 gens
135         // this for loop is just to mutate the protein
136         // Only going thru 1 mutation process right now - Feb 17
137         // Might need to stick with a lower number of mutations,
138         // around 200, 1000 giving me errors and idk why...
139         // Also mutating a protein under the newly proposed
140         // parameter values with this loop
141         for (int h = 0; h < num_mutations; h++) {
142             std::string mutated_protein2 = mutateSeqExpBG(simulated_protein2, new_mut_rate,
new_ind_rate);
143             simulated_protein2 = mutated_protein2;
144         }
145
146         //Adding entropy of sequence to vector for plotting
147         double entropy = calc_entropy(simulated_protein2);
148         entropy_vec[k] = entropy;
149         // 3. Next we will get the average of 10 vectors of summary statistics
150         // For newly proposed parameter values
151
152         std::vector<double> sim_prot_vtr_2 = sim_protein(simulated_protein2);
153         vec_of_vecs2.push_back(sim_prot_vtr_2);
154
155         // Getting the 10 distances in case we need t-test
156         std::vector<double> vec_normal = normalize_vector(sim_prot_vtr_2);
157         double length = sim_prot_vtr_2[0];
158         double numlcr = sim_prot_vtr_2[1];
159         double avgent = sim_prot_vtr_2[2];
160         double length2 = sim_prot_vtr_2[0];
161         length_vec2.push_back(length2);
162         length_vec.push_back(length);
163         num_lcrs_vec.push_back(numlcr);
164         avg_entropy_vec.push_back(avgent);

```

```

165     double dist = vectors_distance2(vec_normal, obs_prot_vtr);
166     distances.push_back(dist); //storing the 10 distances to get mean and stdev
167     distances_ttest[k] = dist; //storing each of the 10 distances for output
168
169 }
170
171 double expected_mutations = Mean(length_vec2)*mutation_rate; //getting the # of expected
mutations
172 std::cout << expected_mutations << "\n";
173 vec_of_vecs3.push_back(distances); //pushing back all 10 distances
174 vec_of_vecs_length.push_back(length_vec); //pushing back lengths
175 vec_of_vecs_avgnum.push_back(num_lcrs_vec);
176 vec_of_vecs_entropy.push_back(avg_entropy_vec);
177
178 // TO TEST THE OUTPUT/PRINT THE VECTOR
179 //for (int x = 0; x < vec_of_vecs.size(); x++) {
180 //    for (int y = 0; y < vec_of_vecs[i].size(); y++){
181 //        std::cout << vec_of_vecs[x][y] << " " << "\n";
182 //    }
183 //}
184
185 // Getting the average vector of the 10 simulated vectors
186 // for newly proposed parameters
187 std::vector<double> sim_prot_vtravg2 = vectors_average(vec_of_vecs2);
188
189 // ADDED STEP - LETS NORMALIZE THE VECTOR
190 sim_prot_vtravg2 = normalize_vector(sim_prot_vtravg2);
191
192 // 5. Lets try finding the distance between the vectors
193 double distance_new = vectors_distance2(sim_prot_vtravg2, obs_prot_vtr);
194
195 current_distances_arr[i] = distance_current; //storing current distances
196 proposed_distances_arr[i] = distance_new; //storing proposed new distances
197
198 // 6. Does the distance satisfy the conditions?
199 // If it does, add new parameter to corresponding list
200 // otherwise stay at same value
201 // Updated the code now to calculate the current state of
202 // the model and the proposed state of the model to compare
203 // those distances.
204
205 if (distance_new < distance_current && new_mut_rate > 0 && new_ind_rate > 0) {
206     mutation_rate = new_mut_rate;

```

```

207         indel_rate = new_ind_rate;
208         mut_rate_arr[i] = mutation_rate;
209         ind_rate_arr[i] = indel_rate;
210         distance_array[i] = distance_new;
211         index[i] = i;
212         accepted_rejected[i] = 1;
213         acceptance_counter = acceptance_counter + 1;
214         acc_rej_rate[i] = acceptance_counter/i;
215         distance_current = distance_new; //setting the new current distance to the new distance so we
can compare a new distance to this accepted one
216         std::cout << "ACCPETED" << "\n" << "\n";
217     } else { //Do a ttest
218         double standard_dev = stdDev(distances);
219         double mean = Mean(distances);
220         double tstat = abs((mean - distance_current) / (standard_dev / sqrt(distances.size())));
221         double p_value = tTest(distances, distance_current);
222
223         //Adding ttest values to output
224         mean_ttest_arr[i] = mean; //storing means from ttest of 10 vectors
225         stddev_ttest_arr[i] = standard_dev; //storing stdev from 10 vectors
226         tstat_array[i] = tstat; //storing test statistics
227         probability_ttest[i] = p_value; //storing pvalues for ttest
228
229         double random_number = myran.doub(); //Random number 0-1
230         if (random_number < p_value && new_mut_rate > 0 && new_ind_rate > 0) {
231             std::cout << "ACCEPTED pval" << "\n" << "\n";
232             mutation_rate = new_mut_rate;
233             indel_rate = new_ind_rate;
234             mut_rate_arr[i] = mutation_rate;
235             ind_rate_arr[i] = indel_rate;
236             distance_array[i] = distance_new;
237             index[i] = i;
238             accepted_rejected[i] = 1;
239             acceptance_counter = acceptance_counter + 1;
240             acc_rej_rate[i] = acceptance_counter/i;
241             distance_current = distance_new;
242         } else {
243             std::cout << p_value << "\n" << random_number << "\n";
244             std::cout << "NOT ACCEPTED pval" << "\n" << "\n";
245             std::cout << i << "\n";
246             mut_rate_arr[i] = mutation_rate;
247             ind_rate_arr[i] = indel_rate;
248             distance_array[i] = distance_current;

```

```

249         index[i] = i;
250         accepted_rejected[i] = 0;
251         acc_rej_rate[i] = acceptance_counter/i;
252         continue;
253     }
254 }
255 }
256
257 std::ofstream myfile("accepted_parameters.txt"); //Create and open txt file
258 // Printing the arrays of parameter values
259 for (int b = 0; b<num_simulations; b++) {
260     myfile << index[b] << '\t' << mut_rate_arr[b] << '\t' << ind_rate_arr[b] << '\t' <<
distance_array[b] << '\t' << '\t' << current_mut_rate_arr[b] << '\t' << current_ind_rate_arr[b] << '\t' <<
261     proposed_mut_rate_arr[b] << '\t' << proposed_ind_rate_arr[b] << '\t' << '\t' <<
mean_ttest_arr[b] << '\t' << stddev_ttest_arr[b] << '\t' << tstat_array[b] << '\t' <<
probability_ttest[b] << '\t'
262     << accepted_rejected[b] << '\t' << '\t' << acc_rej_rate[b] << '\n';
263 }
264
265 std::ofstream myfile2("10_distances.txt");
266 for (int x = 0; x<vec_of_vecs3.size(); x++) {
267     for (int y = 0; y<vec_of_vecs3[x].size(); y++) {
268         myfile2 << index[x] << '\t' << '\t' << proposed_mut_rate_arr[x] << '\t' <<
proposed_ind_rate_arr[x] << '\t' << vec_of_vecs3[x][y] << '\t' << vec_of_vecs_length[x][y]
269         << '\t' << vec_of_vecs_avgnum[x][y] << '\t' << vec_of_vecs_entropy[x][y] << '\t' <<
entropy_vec[y] << '\n';
270     }
271 }
272 }

```

Appendix 3: Simulated Protein Summary Statistics

/home/alex/scratch/abcmcmc-thesis4c12/abcmcmc.cpp/simulated_protein.cpp

```

1 #include <iostream>
2 #include <fstream>
3 #include <string>
4 #include <vector>
5
6 ///////////////////////////////////////////////////
7 // SECOND FUNCTION READS IN SRP40 PROTEIN AND CREATES A VECTOR WITH/
8 // CHARACTERISTICS - THIS IS EITHER S or D - THE ORIGINAL DATA      /
9 ///////////////////////////////////////////////////
10
11 std::vector<double> sim_protein(std::string sim_prot){
12
13     std::vector<double> sim_ss_vec; //Making the simulated summary statistic vector
14
15     // Finding the LENGTH of protein sequence
16     int length = sim_prot.length();
17     sim_ss_vec.push_back(length);
18
19     // Creating fasta file of simulated protein string to use in segA
20     std::ofstream MyFile("simulated_protein.fasta"); //Create and open a fasta file
21     MyFile << ">simulation\n" << sim_prot; //Write to the file
22     MyFile.close(); //Close the file
23
24     // Finding the number of LCRs
25     system("segA simulated_protein.fasta 15 1.9 2.2 -1 | grep '>' | wc -l >> numlcr_sim.txt"); //Write
26     the results of counting the LCRs to a txt file
27     std::ifstream myreadfile3("numlcr_sim.txt"); //Declaring fstream variable and connecting it to a
28     stream object by opening the file
29     double numlcr_sim; //Declare int variable which holds the contents of the 'numlcr.txt' file
30     myreadfile3 >> numlcr_sim; //Pipe file's content into stream
31     sim_ss_vec.push_back(numlcr_sim); // using push_back to append to vector
32
33     //This if statement is to remove the division by zero error in
34     //the awk command below
35     if (numlcr_sim == 0) {
36         sim_ss_vec.push_back(0);
37         //Removing files so they are empty after
38         system("rm numlcr_sim.txt");
39         system("rm simulated_protein.fasta");

```

```

38     } else {
39
40         // Finding the average entropy of the LCRs dont know why I have
41         // to use egrep for the extraction of numbers part in the system
42         // call
43         system("segA simulated_protein.fasta 15 1.9 2.2 -l | grep 'complexity' | awk '{print $2}' | egrep -o
44             '([0.0-9.0]+)' | awk '{sum += $1} END {print mean=sum/NR}' >> avg_entropy_lcrs_sim.txt");
45         std::ifstream myreadfile4("avg_entropy_lcrs_sim.txt");
46         double avgentropylcrs_sim;
47         myreadfile4 >> avgentropylcrs_sim;
48         sim_ss_vec.push_back(avgentropylcrs_sim);
49         //Removing files so they are empty after
50         system("rm avg_entropy_lcrs_sim.txt");
51         system("rm numlcr_sim.txt");
52         system("rm simulated_protein.fasta");
53     }
54
55     // TO TEST THE OUTPUT/PRINT THE VECTOR
56     /*for (int x = 0; x < sim_ss_vec.size(); x++) {
57         std::cout << sim_ss_vec[x] << ' ' << "\n";
58     }*/
59     return sim_ss_vec;
60 }
61
62 //int main() {
63
64     //sim_protein("
65     MASKKIKVDEVPKLSVKEKEIEEKSSSSSSSSSSSSSSSSSSSSSSSGESSSSSSSSSSSSSDSSDSSDSESSSSSSSSSSSSSDSESSSESDSSSSGSSSSSSSS
66     ");
67     //    sim_protein(createSeq(400));
68     //    return 0;
69     //}

```


Appendix 4: Euclidean Distance and Normlization of Vectors

/home/alex/scratch/abcmcmc-thesis4c12/abcmcmc.cpp/distance2.cpp

```

1 #include <iostream>
2 #include <string>
3 #include <iterator>
4 #include <algorithm>
5 #include <vector>
6 #include <functional>
7 #include <cmath>
8 #include <numeric>
9
10 //this function I made to normalize vectors
11 //FEB 23 - THIS IS NOT HOW IM GOING TO NORMALIZE
12 /*std::vector<double> normalize_vector(std::vector<double> vec1) {
13     std::vector<double> normalized_vec;
14     double normalize_value = 0.00;
15     int vecSize = vec1.size();
16     for (int i = 0; i < vecSize; i++) {
17         normalize_value += vec1[i] * vec1[i];
18     }
19     for (int j = 0; j < vecSize; j++) {
20         normalized_vec.push_back(vec1[j]/sqrt(normalize_value));
21     }
22
23     // TO TEST THE OUTPUT/PRINT THE VECTOR
24     for (int x = 0; x < normalized_vec.size(); x++) {
25         std::cout << normalized_vec[x] << ' ' << "\n";
26     }
27
28
29     return normalized_vec;
30 }*/
31
32 //Feb 23 - New function for normalizing
33 std::vector<double> normalize_vector(std::vector<double> vec1) {
34     std::vector<double> normalized_vec;
35     int srp40_length = 406;
36     int vecSize = vec1.size();
37     float length = vec1[0];
38     float num_lcrs = vec1[1];
39     double avg_entropy = vec1[2];

```

```

40
41 //setting the max and mins
42 float max_length = srp40_length*1.5;
43 float min_length = srp40_length*0.5;
44
45 double max_lcrs = srp40_length/8; //8 is the hypothetical length of the shortest LCR, this will give
you a maximal total number of LCRs
46 double max_entropy = 4.3;
47
48 //pushing back the normalized values
49 normalized_vec.push_back((length - srp40_length + min_length)/srp40_length);
50 normalized_vec.push_back(num_lcrs/max_lcrs);
51 normalized_vec.push_back(avg_entropy/max_entropy);
52
53 // TO TEST THE OUTPUT/PRINT THE VECTOR
54 /*for (int x = 0; x < normalized_vec.size(); x++) {
55     std::cout << normalized_vec[x] << ' ' << "\n";
56 }*/
57
58 return normalized_vec;
59
60 }
61
62 //this is my function I made to solve euclidean distance
63 double vectors_distance2(std::vector<double> a, std::vector<double> b) {
64     int vecSize = a.size();
65     double dist = 0.0;
66     for (int i = 0; i < vecSize; i++){
67         dist += pow((a[i] - b[i]), 2);
68     }
69
70     //std::cout << sqrt(dist) << "\n";
71     return sqrt(dist);
72 }
73
74 //FOR TESTING THIS FILE
75 /*int main() {
76     std::vector<double> v1 = {500,6,2.678};
77     normalize_vector(v1);
78 }*/

```

Appendix 5: One Sample t-test

/home/alex/scratch/abcmcmc-thesis4cl2/abcmcmc.cpp/ttest.cpp

```

1 #include <iostream>
2 #include <bits/stdc++.h>
3 #include <vector>
4 #include <iomanip>
5 using namespace std;
6
7 //Function to find Mean
8 double Mean(std::vector<double> vec) {
9     double sum = 0;
10    for (int i=0; i < vec.size(); i++) {
11        sum = sum + vec[i];
12    }
13
14    return sum/vec.size();
15 }
16
17 //Function to calculate Standard deviation of given vector
18 double stdDev(std::vector<double> vec) {
19     double sum = 0;
20     for (int i = 0; i < vec.size(); i++) {
21         sum = sum + (vec[i] - Mean(vec)) * (vec[i] - Mean(vec));
22     }
23
24     return sqrt(sum / (vec.size() - 1));
25 }
26
27 //Function to get the equation of a line and estimate the
28 //probability based on the t-statistic
29 double eqn_line(double y2, double y1, double x2, double x1, double tstat) {
30     double m = (y2 - y1) / (x2 - x1);
31     double b = y2 - (m*x2);
32     double x = (tstat - b) / m;
33
34     return x;
35 }
36
37 //ReWriting function to obtain estimate of pvalue from linear
38 //interpolation
39 double tTest(std::vector<double> vec1, double prev_dist) {

```

```

40     double mean1 = Mean(vec1);
41     double sd1 = stdDev(vec1);
42
43     //Equation to get t-statistic for 1 sample
44     double t_statistic = abs((mean1 - prev_dist) / (sd1 / sqrt(vec1.size())));
45
46     std::vector<double> areas = {0.5, 0.25, 0.2, 0.15, 0.10, 0.05, 0.025, 0.01, 0.005, 0.001, 0.0005}; //
    this is like y
47     //std::vector<double> crit_t_values = {0.000, 0.703, 0.883, 1.100, 1.383, 1.833, 2.262, 2.821, 3.250,
    4.297, 4.781}; //critical t values 9 degrees of freedom - this is like x
48     //std::vector<double> crit_t_values = {0.000, 0.677, 0.845, 1.042, 1.290, 1.660, 1.984, 2.364, 2.626,
    3.174, 3.390}; //critical t values 99 degrees of freedom
49     std::vector<double>crit_t_values = {0.000, 0.675, 0.842, 1.037, 1.282, 1.646, 1.962, 2.330, 2.581,
    3.098, 3.300}; //critial t values 1000 degrees of freedom
50
51     int minPosition=0;
52     for (int i = 0; i < areas.size(); i++) {
53         if (t_statistic > crit_t_values[i]) {
54             minPosition = i;
55         }
56     }
57     int maxPosition = minPosition + 1;
58
59     double pval_est = eqn_line(crit_t_values[minPosition], crit_t_values[maxPosition], areas[minPosition
    ], areas[maxPosition], t_statistic);
60
61     return pval_est;
62     return minPosition;
63     return abs(t_statistic);
64 }
65
66
67 //Function to find p value from table above
68 //Since im doing average of 10, the degrees of freedom would be 10-1
69 //which is 9 - Using 9 degrees of freedom all the time so crit
70 //values will be the same all the time
71 //Larger P value means not statically significant.
72
73 //To test the functions
74 /*int main() {
75     std::vector<double> testvec1 = {10,20,30,40,50,60,70,80,90,100};
76     double old_dist = 27.00;
77

```

```
78     std::cout << tTest(testvec1, old_dist);  
79  
80     return 0;  
81 }*/
```