A DATA MINING STUDY OF G-QUADRUPLEXES AND THEIR EFFECT ON DNA REPLICATION

A THESIS IN

Computer Science

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

MASTER OF SCIENCE

by

GREGORY SHANNON NICHOLS

B.S., University of Missouri-Kansas City, 1991
MBA, Bellevue University, 2003

Kansas City, Missouri
2012

A DATA MINING STUDY OF G-QUADRUPLEXES AND

THEIR EFFECT ON DNA REPLICATION

Gregory Shannon Nichols, Candidate for the Master of Science Degree

University of Missouri-Kansas City, 2012

ABSTRACT

G-quadruplexes are guanine rich sequences of DNA that can form non-Watson-Crick four stranded structures. They have been found to exist in various regions of the genome and are believed to play a biological role. We hypothesize that the presence of these structures poses a barrier to DNA replication by standard DNA polymerases and thus requires the intervention of alternative robust but error-prone polymerases for the completion of DNA replication. To test this hypothesis *in silico*, we assumed that the presence of error-prone replication could be inferred by studying the degree of variation at these sites. We analyzed the density of single nucleotide polymorphisms in the neighborhood of potential G-quadruplex sequences in the human genome. The analysis shows a significantly higher density of single nucleotide polymorphisms within G-quadruplexes. Further, there is evidence of a directional bias in the extent of error, seen as an asymmetry in the incidence of single nucleotide polymorphisms on either side of quadruplexes. Taken together, the evidence favors the hypothesis that G-quadruplexes have a deleterious effect on the fidelity of DNA replication.

A secondary research goal of the thesis is to reduce the number of false positives in the prediction of G-quadruplexes based only on sequence information. Most current

algorithms are regular expression searches based on sequences that have shown potential to form G-quadruplexes. Using the results from our investigation on sequence variation, predicted melting temperature and machine learning models, attributes derived solely from the sequences were analyzed to determine if classification can be accurately performed. We conclude that factors external to the sequence may be important in determining if and when G-quadruplexes form.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Computing and Engineering, have examined a thesis titled “A Data Mining Study of G-quadruplexes and Their Effect on DNA Replication”, presented by Gregory Shannon Nichols, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Deendayal Dinakarpandian, M.D, Ph.D, M.S., Committee Chair

Department of Computer Science Electrical Engineering

Praveen R. Rao, Ph.D

Department of Computer Science Electrical Engineering

Yugyung Lee, Ph.D

Department of Computer Science Electrical Engineering

TABLE OF CONTENTS

LIST OF EQUATIONS

Equation

LIST OF ILLUSTRATIONS

LIST OF TABLES

Table

CHAPTER 1

INTRODUCTION

1.1 Biology Background

Although this is a thesis for computer science, this chapter provides a small amount of biology background which should prove helpful in following the rationale used throughout. Tozeren and Byers provide significantly more biological details [1].

1.1.1 DNA Replication

Human DNA (deoxyribonucleic acid) is a double-stranded molecule responsible for encoding genetic information within a person. It is composed of the nucleotides adenine (A), guanine (G), cytosine (C), and thymine (T) held together by weak bonds in pairs. Normally pairs are formed between adenine and thymine or guanine and cytosine [1]. Figure 1 shows a typical representation of the double helix DNA strand [2].

DNA replication is the process of synthesizing new DNA strands using existing DNA as the template. Although the process of replication involves the complex interaction of many molecules, a simplified explanation is given here for background use only. Figure 2 gives a typical representation of DNA replication [3].

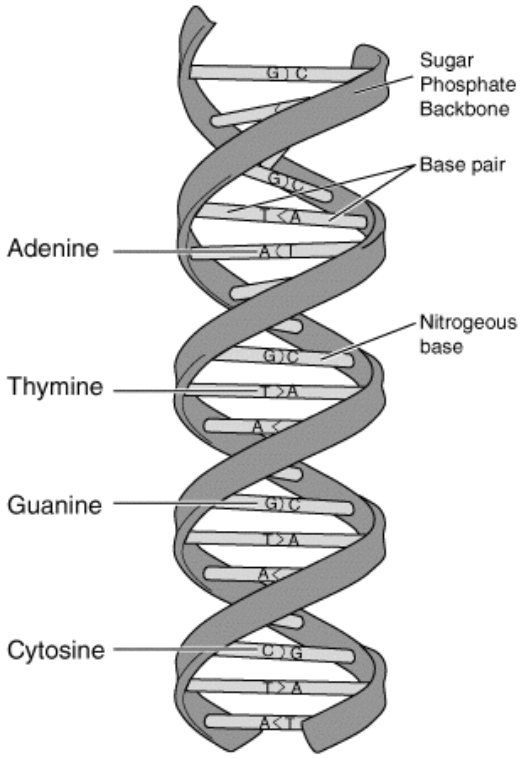

Figure 1 – Typical representation of double helix DNA
Obtained from rinf.com

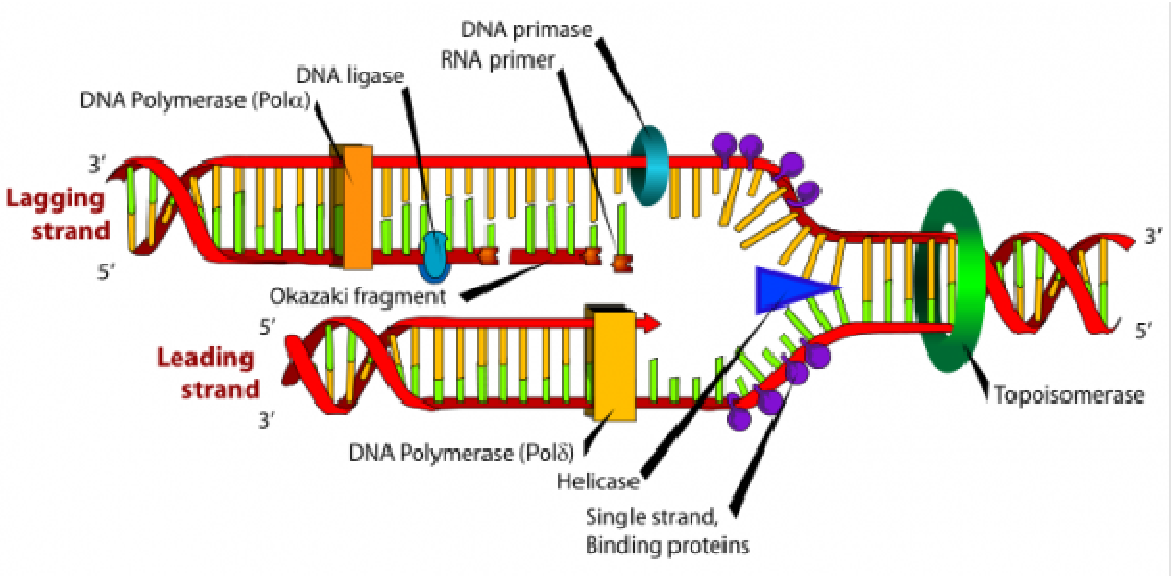

Figure 2 – Simple graphical representation of DNA replication
Obtained from faculty.ksu.edu.sa

Synthesis begins at regions known as replication origins, or replicons, located throughout the DNA. It is estimated that human DNA contains between ten thousand and one hundred thousand replicons. At the replicon, various molecules prepare the DNA for replication by unwinding the molecule and attaching primers. DNA polymerase, itself made of several subunits, attaches to the DNA and is largely responsible for the copying of the template. After the DNA polymerase copies the individual strands and various error checks are performed, the result is two pairs of strands, with each pair having one of the original strands. [4]

1.1.2 G-quadruplexes

Certain DNA sequences rich in guanine (G) have been shown to self-assemble into structures called G-quadruplexes [5][6]. Typically, although there are variations, G-quadruplexes are formed from sequences consisting of four sections containing three guanines each linked by other bases which can also include guanine. At the center of the tetrads is a monovalent cation, such as K^+, that adds further stability. A typical representation of a G-quadruplex is shown in Figure 3 for the sequence UGGGCAGGGCAGGGUGGG (U can represent any base). Note that the cation is not shown. A more detailed explanation of G-quadruplexes and their structure is given by Balasubramanian et al [7].

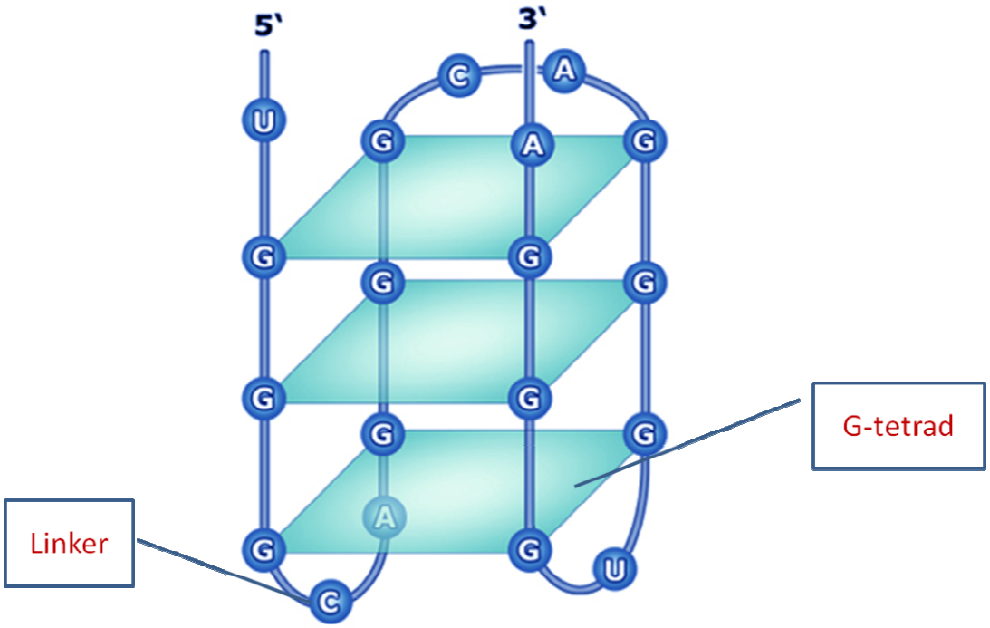

Figure 3 – Typical representation of a G-quadruplex

1.1.3 Single Nucleotide Polymorphism (SNP)

Single-nucleotide polymorphisms (SNPs) are variations within the genome of a biological species where members have different nucleotide bases at a particular location. An example can be seen in the two DNA segments, AATTGCA and ACTTGCA. The two different bases in the second position indicate a SNP. Although insertions and deletions are commonly referred to as SNPs, this paper limits its discussion of SNPs to single bases having multiple alleles such as the previous example.

There are many possible reasons for SNPs. One possible origin of a subset of them is due to error during replication. Although the DNA replication process contains an error checking mechanism, not all errors are corrected. This potential type of error plays a central part in the thesis's analysis.

1.1.4 Promoter and Telomere Regions of DNA

Transcription is the process where an RNA (ribonucleic acid) molecule is produced from a sequence of DNA known as a gene. This is the first part of gene expression. The region where RNA polymerase attaches to the DNA and initiates transcription is known as the promoter region.

Another important part of the chromosome (the complete DNA molecule) is the telomere, which is located at the ends of the chromosome. Although the telomere is not directly used for gene expression, one of its purposes is for stability of the DNA molecule.

1.2 Hypothesis

The postulation has been made that the structure of G-quadruplexes may inhibit the proper operation of DNA polymerase during DNA replication; therefore, alternative robust but error-prone polymerase must be used to continue replication [8]. Research on telomere regions has indicated that a separate error-prone telomerase complex is required to replicate the regions containing G-quadruplexes [9][10].

1.2.1 Objective 1

Our first objective is to determine whether the error rate is in fact significantly greater near G-quadruplexes. Currently no technique exists to directly view replication as it progresses through a G-quadruplex or directly measure the error rate. Indirect methods must be employed to test this hypothesis. As noted in the section on SNPs, one possible source of SNPs is due to error during replication. By analyzing the SNP density of regions containing G-quadruplexes, we hope to accomplish our stated objective.

1.2.2 Objective 2

As a second objective, we wish to use the information gained from analyzing SNP density to help reduce the number of false positives returned during the prediction of G-quadruplexes solely from sequence information. Current methods are limited to regular expression matching and are known to contain many false positives. Reducing this false positive rate would be extremely helpful in analyzing G-quadruplexes.

CHAPTER 2

SNP DENSITY ANALYSIS

2.1 Methods

2.1.1 PQS Generation

A list of G-quadruplexes must be obtained before any analysis can be performed. Although many sequences have been shown to form G-quadruplexes *in vitro* under appropriate environmental conditions, no way has been discovered to definitively determine what sequences will form G-quadruplexes in vivo. Many different methods have been developed in an attempt to predict sequences with the potential of becoming G-quadruplexes [11]-[13]. These sequences have been given the name; putative quadruplex sequences (PQS).

The QuadParser program, using default settings, was used with NCBI build 37 of the human genome to generate a list of PQS for use in this thesis. QuadParser searches for sequences of the form $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ on the '+' strand and $C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}$ on the '-' strand. Huppert and Balasubramanian have provided data supporting that the PQS obtained from QuadParser can form G-quadruplexes [6].
After using QuadParser' we had a file for each chromosome listing the beginning and ending genomic coordinates for each PQS, an entry indicating number of overlapping PQS (as discussed below) and the actual sequence.

Overlapping PQS: During the execution of 'QuadParser,' overlapping PQS are shown as a single PQS. Overlapping PQS are instances when the sequence in question

contains multiple potential PQS. Which of the overlapping PQS, if any, actually become a G-quadruplex is not known. Some examples of this overlapping are shown in Table 1.

Table 1 – Example of overlapping PQS

GGGACGGGTGGGCGATGCGGG	1 PQS
GGGGTCCGACGGGACAGGGATAGGGATCGGGG	2 overlapping PQS
GGGTGGGTGGGTGGGTGGGTGGGTGGGTGGG	5 overlapping or 2 non-overlapping PQSs

We obtained a total of 198,520 PQS from QuadParser . Of these, 59,362 had potential overlaps.

2.1.2 PQS vs. Background

When asking whether G-quadruplexes cause error in DNA replication, as demonstrated by SNPs, we should first look at whether there seems to be a larger percentage of SNPs near G-quadruplexes, using PQS as proxies, compared to background. Our first analysis investigates whether the SNP density immediately surrounding a PQS is greater than background SNP density

During this particular analysis, overlapping PQS are treated as single PQS. We do not know which, if any, of the overlapping PQS actually form G-quaduplexes. If we separate the PQS into individual PQS we will have many intervals that differ by only a few index numbers. The overall average we are attempting to find should not be affected by simply treating each overlapping PQS as a single PQS.

For each PQS, three intervals of length 200 bases were created (shown in figure 4). One interval (‘center’) was centered on the midpoint of the PQS. A second interval (‘flanking 1’) was started 700 bases upstream of the beginning of the PQS interval and the third (‘flanking 2’) was created 500 bases downstream of the end of the PQS interval.

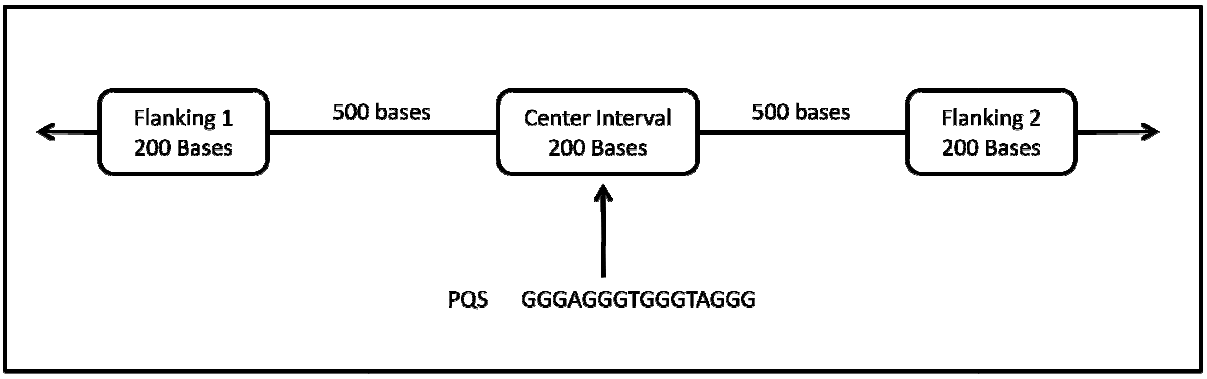

Figure 4 – Illustration of interval creation

The terms upstream and downstream are relative to the genomic coordinates of the PQS. Downstream indicates a smaller index, while upstream indicates a larger index. Whether the PQS originated on the ‘+’ or ‘-‘ strand was ignored. Since we are just comparing the center interval against flanking intervals, the relative position is all that is of concern. No attempt was made to filter the flanking intervals for content. (e.g., they could be located in other PQS, introns, exons, promoter regions, etc.)

Since we do not know the periodicity of the polymerases involved, the choice of interval size was initially picked to ensure the interval size would be large enough to include any SNPs that would be caused during replication of a G-quadruplex, but not so large as to contain a large amount of noise in the form of background SNP density. To narrow the interval size, the location of where each SNP was located within each interval was noted.

The total number of SNPs at each base location was then plotted, resulting in the plots shown in figure 5 which include a smoothing regression line.

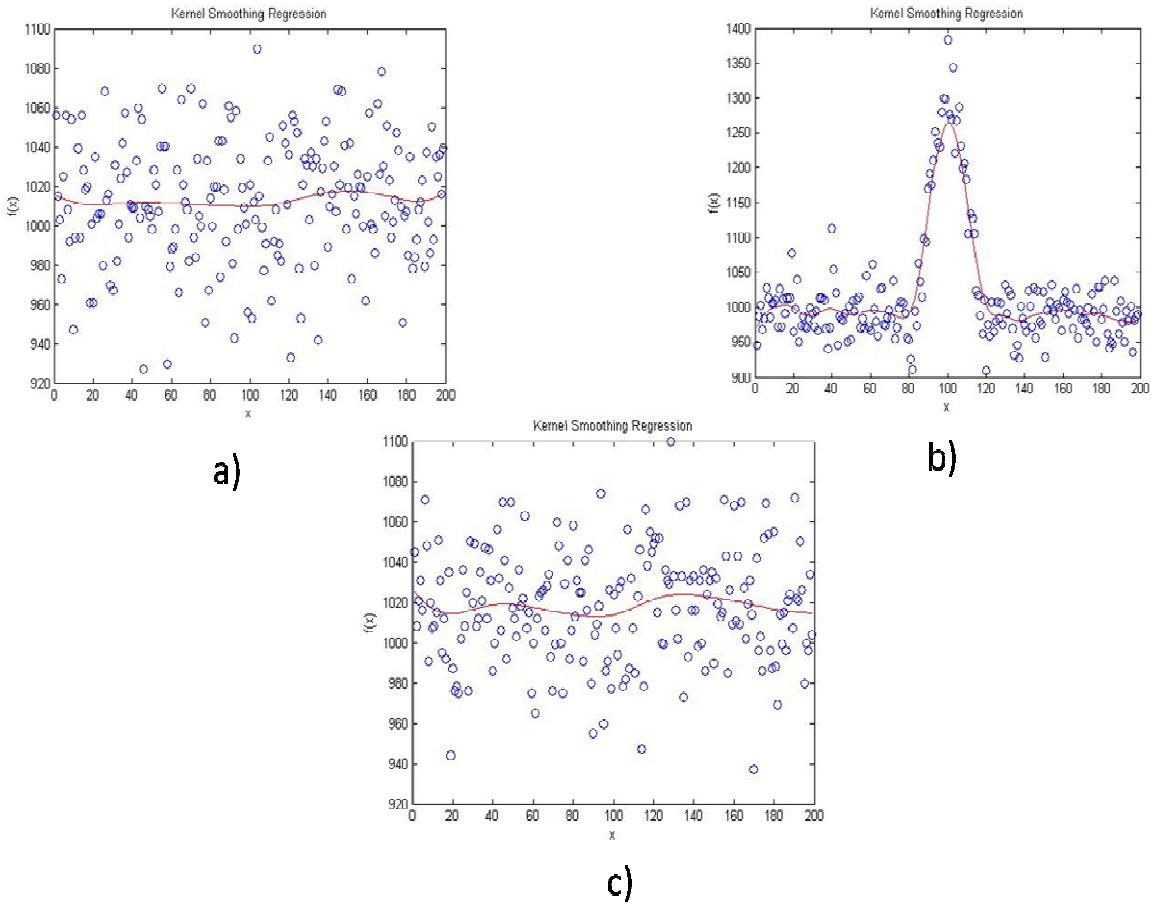

Figure 5 – Plots of SNP counts versus base locations with kernel smoothing regression line. a) Flanking 1 interval; b) Center interval; c) Flanking 2 interval

Local linear kernel regression, using a Gaussian kernel, was used to obtain the regression line. The underlying principle of local linear kernel regression is to fit a straight line locally using the values within a certain distance of the target data point. The Gaussian kernel provides weighting to data points surrounding the target value based on a Gaussian function. The closer the data point to the target, the more weight; therefore they have more

influence on the derived local line. The ‘smoothness’ of the line can be adjusted based on the number of bins that the data is divided into and also on a parameter called the bandwidth, which adjusts the weighting the kernel gives non-target values as the distance increases from the target. A detailed explanation is given by Bowman and Azzalini [25].

Inspection of the flanking interval plots indicates that the background SNP density is approximately 1000 SNPs per base location. The center interval shows this same density except in the 40 to 50 base region near the center, which is the center of the PQS. In an attempt to minimize the amount of noise associated with the background SNP density we decided to focus on the 100 bases centered on the PQS.

Using the web based bioinformatics tool Galaxy, an inner join was performed on the genomic level between the created intervals and known SNP’s from the same build (Build 37) [14][15]. SNP strandedness (‘+’ or ‘-‘) as indicated in the Galaxy database was ignored since a SNP on one strand should indicate a SNP on the other. The number of SNP’s in each one hundred base interval (100 base intervals) was calculated producing a file containing the number of SNP’s for all three PQS intervals.

As noted previously, the flanking intervals were chosen without consideration of location within the genome. It is possible that regions containing SNP density higher than background (e.g. other PQS) are included. In an attempt to minimize this potential noise and reduce the number of statistical tests necessary, the data from the flanking intervals was averaged into one value.

A paired t-test was then performed on the resulting data with the null hypothesis being that the mean difference between each interval is zero. The alternative hypothesis is that the mean for the center interval is greater than the flanking interval mean.

2.1.3 Analysis of Promoter Region PQS

Our first experiment investigated whether SNP density near PQS is greater than background. This just answers the question of whether more SNPs tend to be located near PQS. If our proposal that G-quadruplexes cause greater numbers of SNPs due to using an error-prone polymerase we should see a higher number of SNPs after the error-prone polymerase takes over replication. Unfortunately, we do not know the direction of DNA replication. Instead we are left with investigating whether the number of SNPs is different statistically between the 'upstream' and 'downstream' sides.

Huppert and Balasubramanian show that PQS density throughout the genome is lower than would be expected by probability alone, while there is an enrichment by a factor of 6.4 in promoter regions compared to elsewhere. It can be argued that the increase indicates a possible increase in the proportion of PQS that actually form G-quadruplexes.

We chose to perform the next experiment in the gene promoter regions due to this possibility of a larger percentage of PQS actually forming G-quadruplexes. This should minimize the amount of noise contributed by non G-quadruplex forming PQS.

Two separate analyses were done in this region. The first separated the overlapped PQS into individual PQS and compared the number of SNPs on each side of the PQS midpoint. The overlapped PQS are separated since we need to find the midpoint of each PQS to compare the SNP numbers. If we left the overlapped PQS as a single PQS we have no basis for picking the midpoint.

The second looks only at PQS that originally only contained a single possible PQS (i.e. no overlaps). This was done principally as a check against the results we got with the

analysis of the overlapped PQS that we separated. If we get results that are inconsistent, we should investigate our procedure of separating the overlaps.

The basic procedure for each analysis consisted of the same methods:

Two intervals were created. One interval (upstream interval) started at the PQS midpoint and headed 'upstream' 100 bases and the other (downstream interval) from the midpoint 'downstream' a hundred bases.

A potential problem is encountered with PQS that have an odd number of bases since the intervals should each contain half the PQS. If the odd base is consistently placed in one interval over the other we could potentially add bias to our calculations. Therefore, whenever an odd length PQS was encountered, the center odd base was randomly placed in one or the other interval.

Another complication is that the actual location of the promoter region for each gene varies, sometimes substantially. A quick scan of literature shows a large percentage of known promoter regions being located within a few thousand base pair upstream (5') of the start of the coding sequence for the gene [26]. In order to keep this experiment relatively simple, but have confidence that a large percentage of promoter region PQS are being examined, we chose to include any PQS within 3k upstream of a known gene. We followed the following steps.

1. Galaxy was used to create intervals for each known gene that starts 3k bases upstream (5') and ends at the beginning of the gene.
2. An inner join was performed using Galaxy on the genomic regions from part 1 and a file containing the appropriate PQS along with their strand identification.

3. The join only compares the genomic coordinates of the two intervals being compared, it does not take into account whether the interval was from the '+' or '-' strand. To keep only PQS that are in our promoter region, this data set was filtered to contain only listings that were both either '+' or '-'.
4. Many of the PQS were listed multiple times with different genes. The file was edited to count each PQS only once.
5. As previously done with the background SNP investigation, the intervals were joined with all single type SNP's from Galaxy. Strandedness was neglected at this stage for same reason as previously stated.
6. The number of SNP's in each interval for each PQS was calculated.

The purpose of this experiment is to determine if there is a difference in SNP density 'before' and 'after' the PQS. If the direction of replication were known this would be a straightforward calculation using a paired t-test. Unfortunately, the potential flip-flop of replication direction could cause inaccurate results with a paired t-test. Figure 6 provides a simple illustration of the 'flip flop' problem.

Figure 6 – Illustration of 'flip-flop' problem. In case a) the result would correctly support the hypothesis of more SNPs after replication of PQS; whereas, case b) would falsely support this idea since the direction of replication is reversed for one of the PQS

Taking the absolute difference between the SNP counts from each side of each PQS allows us to ignore this direction affect. The question then becomes how to test whether the results we obtain significantly support our hypothesis. We need the appropriate distribution to compare our results against.

We chose to use a model based on the binomial distribution. Using the absolute differences, the observed distribution was compared to a binomial distribution using the chi-squared goodness-of-fit test.

The expected binomial distribution was determined as follows. Each PQS and its associated intervals were treated as a separate experiment. The total number of SNPs from both intervals was used as the number of trials. A success was considered a SNP in the 'upstream' interval. The probability used was 0.5 which equates to an equal probability for the SNP to be in the 'upstream' or 'downstream' interval.

If the null hypothesis that the SNP density before and after a PQS is the same, then the probability that a SNP is in the 'upstream' interval versus the 'downstream' interval should be the same and the distribution of all the counts should be significantly the same as a binomial distribution with probability of 0.5.

Three bins for each PQS were created. One bin held the observed absolute values, another the expected values based on the binomial model, and the third held the resulting counts from a simulation based on the binomial model.

For each PQS the following procedure was followed to fill the bins. Using the actual SNP counts, the bin for observed value corresponding to the absolute value of their difference was incremented by one. Using the binomial probability equation, the expected values were calculated for all possibilities and these were converted to absolute values and

then the associated expected value was incremented by this probability. The total of all expected values equal one. In addition to the observed and expected values, a simulated value was calculated using the *numpy* binomial distribution for python. This value was placed in the corresponding simulated bin.

The simulated set was calculated to test against the expected value to show the correctness of the chi-squared goodness-of-fit for the binomial distribution.

Table 2 gives an example of the followed procedure were there are 3 SNPs in the 'upstream' interval and one in the 'downstream' interval, resulting in the observed value of two being incremented by one.

This procedure was performed for all PQS in the data set resulting in totals for observed and expected values.

Table 2 - Example of calculation for expected values from case where a total of four SNPs are seen in the two intervals. The corresponding expected absolute difference values are incremented by the associated probabilities**.**

Success	Failure	Probability	Corresponding absolute difference
0	4	0.0625	4
1	3	0.25	2
2	2	0.375	0
3	1	0.25	2
4	0	0.625	4

2.1.4 Analysis of Telomere Region

In an attempt to directly measure 'before' and 'after' values, we investigated PQS located relatively close to the telomere regions of each chromosome. The assumption is that replication will be "heading" toward the telomere at that point. If this assumption is valid, we would no longer be restricted to examining only the difference, but could quantify SNP density before and after replication of the PQS.

The distance from the telomeres that we can assume directionality is based upon the periodicity of the DNA polymerase used in the replication of DNA. As discussed previously, we do know this value, so in an attempt to have enough data points, yet stay within the directionality constraint we chose to analyze all PQS within 500,000 bases of each chromosome end.

The procedure followed is similar to that used for the promoter region. It is as follows.

1. All PQS within 500k bases of the beginning of a chromosome were collected and placed into a file. All PQS within 500k bases of the end of each chromosome were also collected and placed into a separate file.
2. Intervals of length 100 bases 'upstream' and 'downstream' were created as discussed previously in the promoter region section.
3. The intervals were joined with all single type SNP's from Galaxy. Strandedness was neglected at this stage for same reason as previously noted.
4. The number of SNP's in each interval for each PQS was calculated.

2.2 Results

2.2.1 PQS vs. Background

As noted in the methods, an average was calculated for the two flanking intervals. This average was then analyzed against the center interval using a paired t-test. The hypothesis being tested was that the difference in means between each interval is zero. The alternative hypothesis is that the center interval has a greater mean. The appropriate p-value to consider is the one-sided p-value. The results are shown in the table 3.

Table 3 - **:** Paired t-test results from PQS versus background regions

	Center Interval	Average Flanking Interval
Mean	0.737	0.713
Variance	0.814	0.379
Observations	142571	142571
p-value (one-tail)	1.35E-16	

2.2.2 Analysis of Promoter Region PQS

Figures 7 and 8 present the totals in bar graphs for the two sets of data. Using these totals, a chi-squared goodness-of-fit test was done to determine if the observed values fit with the binomial distribution. Tables 4 and 5 present those results.

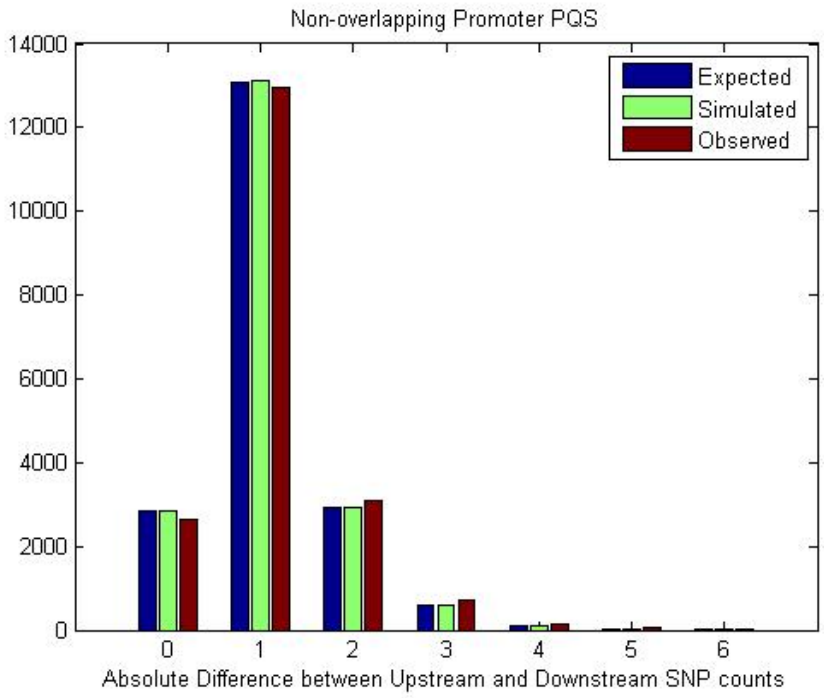

Figure 7 – Distribution of expected, simulated, and observed counts

Table 4 – Results from chi-squared calculation of non-overlapping PQS

Observed vs. Expected Values:		
	Chi-squared number	123.06
	p-value	3.70E-24
Simulated vs. Expected Values		
	Chi-squared number	2.67
	p-value	0.849

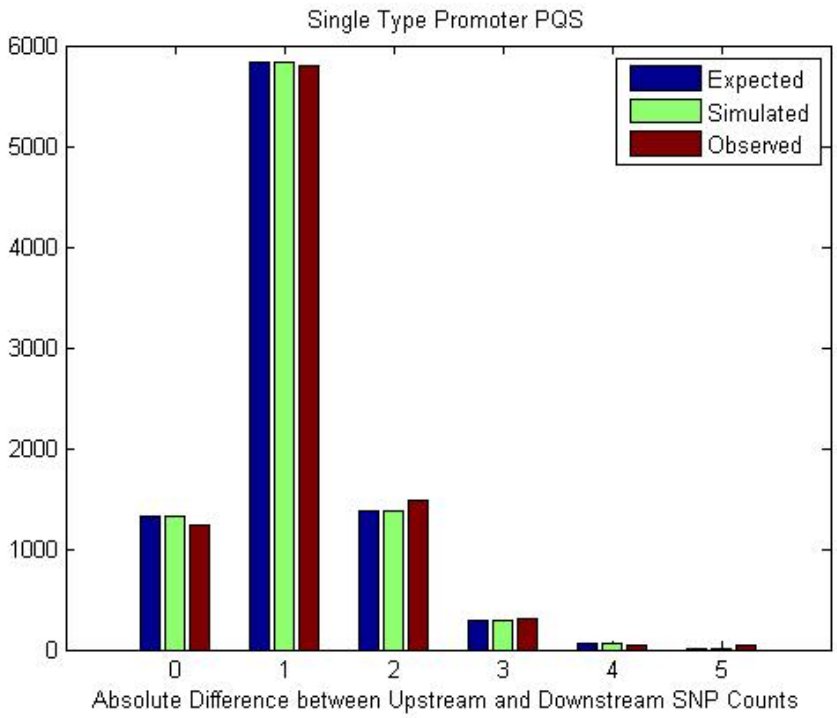

Figure 8 – Distribution of expected, simulated, and observed counts

Table 5 – Results from chi-squared calculation of single-type PQS

Observed vs. Expected Values		
	Chi-squared number	56.55
	p-value	6.25E-11
Simulated vs. Expected Values		
	Chi-squared number	3.22
	p-value	0.67

2.2.3 Analysis of Telomere Region PQS

The SNP values from each PQS were analyzed using a paired t-test. The hypothesis being that the difference in means equals zero. The alternative hypothesis being that the mean for the low index interval SNP count for the *p*-arm telomere is greater than the high index interval SNP count. For the *q*-arm telomere the alternative hypothesis is that the high index interval SNP count is greater than the low index interval SNP count. Tables 6 and 7 present the results.

Table 6 – Results from *q*-arm telomere region

	Low-end index interval	High-end index interval
Mean	0.98	0.95
Variance	1.02	1.12
Observations	2344	2344
p-value (one-tailed)	0.19	

Table 7 – Results from *p*-arm telomere region

	Low-end index interval	High-end index interval
Mean	1.09	0.92
Variance	1.36	0.92
Observations	802	802
p-value (one-tailed)	7.54E-4	

2.3 Discussion

2.3.1 PQS vs. Background

The results show that the regions near PQS have a significantly larger SNP mean than background regions. This result does support the idea that G-quadruplexes may cause greater number of SNPs due to error-prone polymerases during DNA replication.

The results also show a larger variance for the center interval compared to the flanking intervals. Several of the center intervals had a large value for SNPs, whereas this was not the case for the flanking intervals. We know that all PQS do not in fact indicate true G-quadruplexes. This large variance can possibly be explained by the idea that those PQS that do form a G-quadruplex have a larger SNP count. This idea leads directly to the second part of this thesis.

The plot of SNP versus index location for the center interval (figure 3, middle plot) shows a peak slightly right of the interval mid-point. A question is whether the peak shows that SNP density is higher near the center of a PQS or is the peak a consequence of the combination of multi-modal distributions.

If the main hypothesis of this paper that G-quardruplexes tend to create SNPs immediately after encountering a G-quaruplex is true, we would expect to see peak SNP density somewhere close to the end of the PQS. As noted previously the direction of DNA replication is not known exactly in the area of each PQS. This would lead us to believe we should have two peaks, one before the combined PQS counts and one after.

Since we do not have any directionality information we have combined all the data into one distribution. This combined distribution data could easily show one modal point when in fact two 'true' modes exist.

We decided to test whether the data at least supports this possibility. A common method for clustering and decomposing a larger mixture is to use the Expectation Maximization (EM) algorithm to obtain maximum likelihood estimates of the parameters for a Gaussian mixture with *k* components from the data.

The Expectation Maximization algorithm is an iterative process that starts by estimating the parameters and then maximizing the estimated likelihood function. This produces new estimated parameters. This process continues until convergence to local minima. A more detailed explanation is given in a tutorial by Tomasi [27].

Using a value of 2 for *k* we obtained the results shown in table 8.

Table 8 – Parameters obtained from decomposition of the center interval plot

	Mean	Standard Deviation
Original distribution	50.51	28.06
Component One	28.68	16.35
Component Two	73.68	17.18

The average length of all PQS, treating overlaps as a single PQS, is 26 bases. The means calculated from the decomposition of the Gaussian distribution show the conjectured peaks would be on each 'side' of the average PQS.

Unfortunately, this analysis is not conclusive since this is only one of several possible underlying distributions. The main utility of this portion of the analysis is to show that the available data can be consistent with the underlying hypothesis we are investigating, although without information regarding directionality of DNA replication we cannot conclusively show this to be the case.

2.3.2 Analysis of Promoter Region PQS

The results from the promoter regions show that the absolute difference values do not follow the binomial distribution model as would be expected if the null hypothesis were true. This indicates there is a significant bias towards one side or the other, although we cannot say which.

The bar graphs show that the observed values are lower than expected for the zero and one columns, but greater than expected for the higher differences, indicating a bias to one interval having a greater number of SNPs than the other. The result was the same for each sub-category of PQS we examined.

As with the center interval from the background versus PQS experiment, the variance for each interval is relatively large. This may be due to the PQS which actually form G-quadruplexes having a larger number of SNPs and the PQS which do not become G-quadruplexes having results similar to background

2.3.3 Analysis of Telomere region PQS

The p-value for the *p*-arm data indicates that we can reject the null hypothesis. The means also indicate that the low index side of the PQS is higher. If the assumption that the direction of DNA replication is from high index to lower index in this region, then we have

support for our initial hypothesis that error-prone polymerases tend to create SNPs during the DNA replication near G-quadruplexes.

Unfortunately, we do not see the same results with the *q*-arm data. The p-value indicates that we cannot reject the null hypothesis. We are left in the position of needing further analysis to make any determination.

CHAPTER 3

MELTING TEMPERATURE ANALYSIS

3.1 Motivation

As noted in the section on PQS generation, there is no definitive method to distinguish sequences that actually become G-quadruplexes without extensive biological testing. Current methods are based upon regular expression searches for sequences that match patterns that have the potential to form G-quadruplexes. These patterns are general in nature and return a large percentage of sequences that do not in fact become G-quaruplexes. A method more specific than simple pattern matching would be extremely helpful in reducing these false positives.

Given the research presented in the first part of this paper showing a statistically higher SNP density in the region near PQS compared to background, the question arises as to whether PQS with higher SNP density are indicative of true G-quadruplexes.

Answering this question should be relatively easy by comparing the SNP count associated with PQS that actually become G-quadruplexes to those that do not. Unfortunately, there have only been a few sequences shown to form G-quadruplexes in vivo at this time. Most research has been on sequences *in vitro*. Therefore, we are left with trying an indirect method in order to answer our question.

A method we have decided on pursuing is measuring the stability of the potential structure formed by the PQS. In order for a G-quadruplex to perform any meaningful biological role, it must be stable in the biological environment, principally at body temperature. A higher melting temperature is indicative of greater stability within the

biological environment of the human body. If the mean predicted melting temperature of PQS with higher SNP density is statistically higher than those from the lower SNP density PQS, this would support the idea that higher SNP density PQS actually form G-quadruplexes. This information would be valuable in developing a more accurate prediction model.

3.2 Predicting Melting Temperature

Experimentally measuring the melting temperature of a G-quadruplex is a straight forward process with current technologies. Unfortunately, as noted previously, the number of known G-quadruplexes is relatively small. Therefore, the data set currently available is too small to obtain a proper distribution of temperatures to use for significance testing.

Stegle et al. (2009) have developed a Bayesian prediction framework based on Gaussian process regression to determine the thermodynamic stability of unmeasured G-quadruplexes from the sequence information alone [17]. The full details of their method can be found in the reference cited.

The on-line web-service, QuadPredict, implements the methods proposed by Stegle et al [18]. The user provides the PQS and desired monovalent cation concentration and the service provides a predicted melting temperature.

3.3 Method

Two datasets were created for this analysis. All PQS with a SNP count greater than five within the hundred base interval centered on the PQS were collected into one set, resulting in two hundred and thirty six PQS. The other set consisted of two hundred and thirty six PQS selected randomly from all PQS having a SNP count of zero.

The two sets were selected with a large difference in SNP count to minimize potential overlap of temperature distributions if a difference does exist.

Both datasets were input into the QuadPredict web-service along with the default value for cation concentration of one hundred mM of K^+. This concentration is relatively common throughout the human body.

The resulting predicted temperatures were compared and a two-sample t-test performed to test for any significant difference. The null hypothesis being that the difference in temperature means for both sets equals zero. The alternative hypothesis being that the mean temperature for the higher SNP count set is greater than the zero SNP count group. A one-sided p-value is appropriate for this case.

3.4 Results and Discussion

The results from the two-sample t-test are shown in Table 9. The resulting p-value indicates that the null hypothesis, that mean melting temperature difference equals zero, cannot be rejected. In fact, the mean predicted melting temperature for PQS with a SNP count greater than five is lower than that for PQS with zero SNP count.

Table 9 – Two-sample t-test on melting temperatures

	0 SNP PQS	> 5 SNP PQS
Mean predicted melting temperature	64.23°C	64.11°C
Variance	58.70	69.51
p-value (one-sided)	0.44	

Although a significantly higher predicted melting temperature for the high SNP count PQS would have supported the idea that having a high SNP count is indicative of the PQS being a true G-quadruplex; we cannot be too surprised with the outcome. The model used for the QuadPredict web-service has only been optimized on a small number of proven G-quadruplex structures. The authors have indicated that the accuracy of the predictions should increase as more structures are proven and the data is incorporated into the model.

CHAPTER 4

MACHINE LEARNING ANALYSIS

4.1 Motivation

The results from the analysis of predicted melting temperature did not provide a definitive process to minimize false positives in determining G-quadruplexes. Therefore other methods were investigated.

Machine learning models have been used successfully in many circumstances to ascertain patterns inherent in data. These patterns can provide valuable information.

We hope that by using several machine learning models that focus on different types of patterns we may derive rules to minimize the false positives.

4.2 Attributes

Eight attributes which are wholly derived from the sequence data of PQS are calculated for each PQS. These attributes are chosen based on research that has shown some correlation of their values with actual G-quadruplex formation. Table 10 lists the eight attributes.

With the exception of the last attribute entropy, the process of deriving these attributes is obvious. Many publications have been written with regard to entropy and information theory starting with a paper by C. E. Shannon, who is often cited as the inventor of information theory [19].

A basic definition of entropy in relation to information theory is a measure of uncertainty associated with a random variable. For our purposes, it gives a measure to the number of possible ways a sequence can possibly form a G-quadruplex. There may be some

correlation between the number of potential ways, and the probability of forming a G-quadruplex. This attribute allows us to account for this possibility. Figure 9 gives a quick illustration of the process used in calculating entropy for our analysis.

Table 10 – List of sequence derived attributes

PQS total length
Number of tetrads
Average linker length
Percent A composition of linker
Percent C composition of linker
Percent G composition of linker
Percent T composition of linker
Entropy value

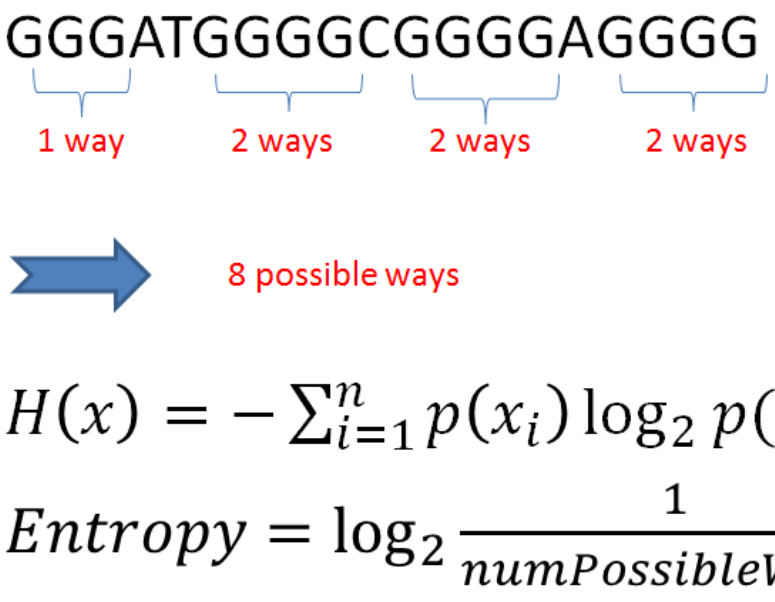

Figure 9 – Steps in calculating entropy value

4.3 Dataset

We have decided to use a high SNP count as indicating a PQS that belongs to the class of G-quadruplexes, and those with a low SNP count not forming G-quadruplexes.

We also chose to perform the analysis on PQS from the promoter region since there has been indication that these may have a higher probability to actually form G-quadruplexes.

If the attributes we are investigating do vary based on whether they belong to the class of G-quadruplexes or not, there is still the possibility that the variation follows some form of distribution. PQS near the boundary of class separation may have overlaps in these values, which would tend to make pattern recognition more difficult. To minimize this possibility, we chose to place PQS with a SNP count of zero in the negative G-quadruplex class and those with four or more in the positive class. This should help highlight any differences there may be.

From the PQS in the promoter region, 11,484 of them had a SNP count of zero. One hundred and two of them had four or more SNPs. To make the classes equal in size, one hundred and two of the zero count PQS were randomly chosen and placed into the class labeled zero. The PQS with four or more were placed in the class labeled one.

For training the subsequent models, seventy percent of the total was placed into the training set. The other thirty percent were used for testing. Both sets were randomly selected with equal proportion of each class represented in each class.

4.4 Models

4.4.1 Naïve Bayesian

For many classification problems, the Naïve Bayesian model is used as a base case for comparison against other models. This model assumes independence between all attributes. Using training data, the probability that the attribute value is seen in each class is calculated. This is done for all attributes.

Since all attributes are considered independent, an example is placed into the class corresponding to the highest probability of having the attribute values associated with the example. This probability is simply the product of all individual probabilities for the attribute values. Given $\boldsymbol{x}(x_1,\ldots,x_d)$, a vector containing the attribute values from an example, the probability that we will have x given a class *C* is given by equation 1[20].

$$\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{C}) = \prod_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{d}} \boldsymbol{p}(\boldsymbol{x_j}|\boldsymbol{C})$$

Equation 1 – Naïve Bayesian equation for probability

4.4.2 Neural Network

Neural network (NN) models were inspired by the biological neuron structure of the brain where many simple elements operate in parallel to perform complicated operations. The basic structure consists of inputs, one or more hidden layers, and the output (Figure 10). The network is trained so that a specific input leads to a specific output such as class. The connection weights between the layers determine the network output. These weights are determined during training by comparing the output of the network with the known class and using any error to adjust the weights.

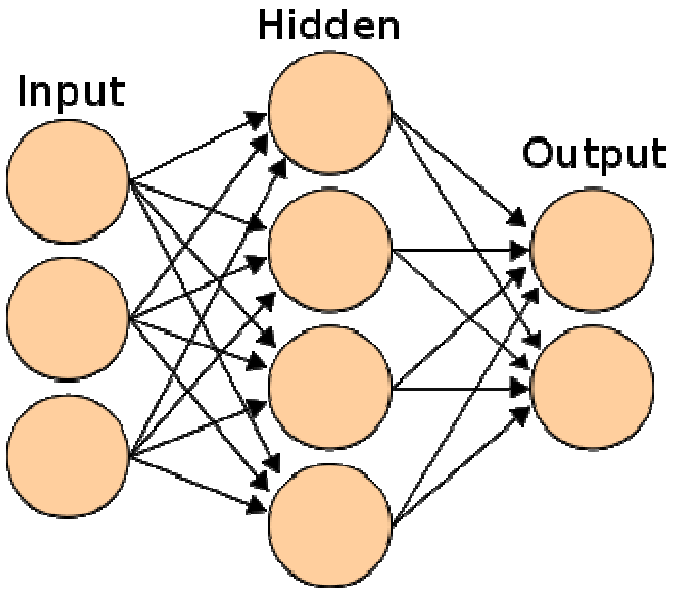

Figure 10 – Graphical representation of neural network

4.4.3 Support Vector Machine (SVM)

The support vector machine model is a relatively recent addition to the regression and classification models available. The idea of SVM can best be visualized with a simple linear two-class data set as shown in figure 11. Two different classes are represented by pluses and time signs. Using the class inputs, a boundary line separating the two classes is found such that the distance to the margin lines is maximized. The margin lines are lines created by passing though a set of class inputs. These inputs are then called supporting vectors.

Unfortunately, most problems are not as easily separated into classes. For these cases, the data is transformed into higher dimensions. Hopefully a hyper-plane analogous to the boundary line can be found in this higher dimension [21].

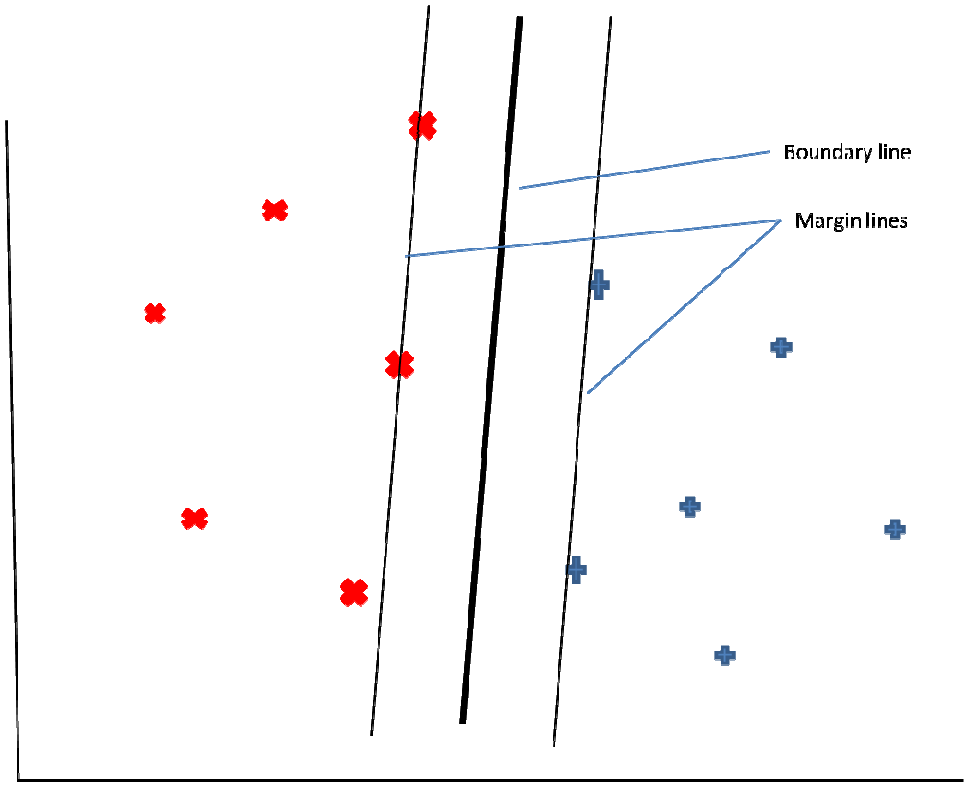

Figure 11 – Graphical representation of support vector machine

4.4.4 Hierarchical Clustering

Although derived attributes can provide useful information in many cases, forcing a model to account for these attributes can in fact lead to false outcomes. In an attempt to counteract this possibility we analyzed the PQS based on sequence information alone. One way to analyze PQSs with only the nominal sequence information (i.e. no derived attributes) is to use some method of clustering limited to sequence information only. One popular basic method of clustering is hierarchical clustering.

This method attempts to cluster items based upon their similarity. Items most similar are placed together in a group [22]. The key to this method is determining a measure for

similarity. For some datasets this is an obvious choice, but others present more challenge since a mathematical construct must be envisioned.

4.4.5 Markov Clustering (mcl)

The results from the hierarchical clustering indicated we should try a more robust form of clustering. We chose Markov Clustering (mcl), an unsupervised cluster algorithm for networks based on simulation of stochastic flow in graphs [23].

Markov clustering is a form of graph clustering where the vertices are the data points to be clustered and the edges are weighted based on some similarity between the points.

The general idea is that vertices will have more edges between vertices in the same cluster than vertices of other clusters. Random walks along the edges will tend to stay within specific clusters. By performing these random walks, we hope to find patterns due to clustering. Markov clustering uses Markov chains to perform these random walks.

4.5 Methods

4.5.1 Naïve Bayesian

The NaiveBayes.fit module of MatLab was used with the training data to create a classifier. Using the resulting classifier, the test data was used to determine the accuracy of the model.

4.5.2 Neural Network

MatLab's neural network pattern recognition tool (nnrtool) was used to create the neural network model. This tool uses scaled conjugate gradient back-propagation to calculate error. The tool also allows the user to choose the number of hidden neurons. The default number of ten neurons returned the best accuracy.

4.5.3 Support Vector Machine (SVM)

The support vector machine module in MatLab allows the user to select various parameters. Based on the paper by Chih-Wei Hsu et al. we chose to use the Gaussian Radial Basis Function kernel (rbf) for the kernel function [24].

With the radial basis function, the user is allowed to choose a scaling factor, sigma, which has a default of one. We varied the value of sigma from one-tenth to four. The default of one returned the best accuracy.

4.5.4 Hierarchical Tree

We chose to use Hamming distance as the measure for similarity**.** Hamming distance between two sequences is a measure of the minimum number of substitutions required to change one string into the other.

We used MatLab's clusterdata module to perform the clustering. For distance, 'hamming' was used and the default linkage value of 'single' was used. The 'single' linkage value finds the shortest distance between clusters.

4.5.5 Markov Clustering (mcl)

The mcl program obtained from www.micans.org was used to perform the Markov clustering. The Needleman-Wunsch module of Biopython was used to perform a global alignment on the sequences. This returned a pairwise similarity score between all sequences. These scores were the input for the mcl program.

4.6 Results

4.6.1 Naïve Bayesian

The model had an accuracy of fifty-five and sixth tenths percent when tested with the test set. The resulting confusion matrix is shown in Table 11. The resulting F-score is 0.44.

Table 11 – Confusion matrix from Naïve Bayesian model

	Class 0	Class 1
Class 0	51	20
Class 1	40	24

4.6.2 Neural Network

With ten hidden neurons the accuracy was fifty-six and one-half per cent. The resulting confusion matrix is presented in Table 12. The resulting F-score is 0.61.

Table 12 – Confusion matrix from Neural Network

	Class 0	Class 1
Class 0	16	8
Class 1	22	23

4.6.3 Support Vector Machine (SVM)

Table 13 presents the confusion matrix, which shows an accuracy of fifty-eight per cent. The resulting F-score is 0.54.

Table 13 – Confusion matrix from Support Vector Machine model

	Class 0	Class 1
Class 0	23	8
Class 1	21	17

4.6.4 Hierarchical Tree

All data points were placed within the same leaf. This indicates that no discernible pattern could be determined.

4.6.5 Markov Clustering (mcl)

The results were consistent with the hierarchical tree analysis. All data points were determined to be in the same cluster.

4.7 Discussion

Solely using sequence information, no discernible patterns or classification was possible with the various machine learning models investigated. By limiting the models to sequence information, external factors were not considered.

For example, cation concentration as well as interaction with various proteins may significantly contribute to the ability of a G-quadruplex to form. Further research may indicate external factors appropriate to include in future models.

CHAPTER 5

CONCLUSION

The first part of our analysis shows significantly the presence of a higher degree of variation in the immediate region surrounding a PQS. This was shown though the presence of a higher SNP density. In addition, asymmetry in the incidence of SNPs on either side of PQS shows a directional bias in the extent of error. These results support our initial hypothesis that the presence of G-quadruplexes pose a barrier to DNA replication by standard DNA polymerases, thus requiring an alternative robust but error-prone polymerase for the completion of DNA replication.

A better biological understanding of G-quadruplexes, particularly the ability to predict sequences that become G-quadruplexes would greatly enhance our ability to corroborate these results.

Attributes derived solely from sequence information did not provide appropriate information to accurately perform classification or derive any patterns to use in the prediction of G-quadruplexes. Biological factors external to the sequence are the most likely reason.

Many simplifying assumptions were made during this research. Future work should attempt to remove some of these assumptions.

Limiting the analysis of promoter PQS to verified promoter regions is one possible improvement. This would require extensive research since at present the current sources do not directly label these regions.

A recent research paper introduces a new database containing 433 human replication-origin sites [27]. Using these and possible future additions could add direct directionality results to our analysis.

REFERENCE LIST

[1] Aydin Tozeren and Stephen W. Byers, "Macromolecules of Life" in *New Biology for Engineers and Computer Scientists,* Upper Saddle River, NJ: Pearson Education, 2004, ch. 2.

[2] DNA-structure-and-bases.png, March 26, 2012, http://en.wikipedia.org/wiki/File:DNA-structure-and-bases.png

[3] DNA-replication, March 23, 2012, http:// faculty.ksu.edu.sa

[4] Aydin Tozeren and Stephen W. Byers, "Cell Division and Its Regulation" in *New Biology for Engineers and Computer Scientists,* Upper Saddle River, NJ: Pearson Education, 2004, ch. 7.

[5] Julian L. Huppert, "Structure, location and interactions of G-quadruplexes," *The FEBS Journal,* vol. 277, pp. 3452–3458, 2010.

[6] Julian L. Huppert and Shankar Balasubramanian, "Prevalence of quadruplexes in the human genome," *Nucleic Acids Research*, vol. 33, no. 9, pp. 2908–2916, May 24, 2005.

[7] Shankar Balasubramanian et al., " Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?," *Nature Reviews*, vol. 10, pp. 261-275, April 2011.

[8] citation for peter

[9] Myron F. Goodman, "Error-Prone Repair DNA Polymerases in Prokaryotes and Eukaryotes," *Annu. Rev. Biochem*, vol. 71, pp. 17-50, January 31, 2002.

[10] Peter Baumann, "Taking control of G-quadruplexes," *Nature Structural and Molecular Biology*, vol. 12, no. 10, pp. 832-833, October 2005.

[11] Alan K. Todd et al., "Highly prevalent putative quadruplex sequence motifs in human DNA," *Nucleic Acids Research,* vol. 33, no. 9, pp. 2901–2907, May 24, 2005.

[12] H. M. Wong et al., "A Toolbox for Predicting G-Quadruplex Formation and Stability," *Journal of Nucleic Acids*, vol. 10, pp. 1-6, 2010

[13] A. K. Todd, "Bioinformatics Approaches to Quadruplex Sequence Location," *Methods*, vol. 43, pp. 246-251, 2007

[14] Belinda Giardine et al., "Galaxy: A platform for interactive large-scale genome analysis," *Genome Research,* vol. 15, pp. 1451-1455, September 2005.

[15] J. Goecks et al.," Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biol* , vol. 86, 2010.

[16] Julian L. Huppert and Shankar Balasubramanian, "G-quadruplexes in promoters throughout the human genome," *Nucleic Acids Research*, vol. 35, pp 406-413, 2007

[17] Oliver Stegle et al., "Predicting and understanding the stability of G-quadruplexes," *Bioinformatics*, vol. 25, pp i374-i382, 2009

[18] H. Wong et al. *Quadruplex.org.* University of Cambridge. Available at http://www.quadruplex.org (Last accessed date March, 2012)

[19] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379-423, July 1948

[20] Ethem Alpaydin, "Graphical Models," in *Introduction to Machine Learning*, 2nd ed. Cambridge, MA: The MIT Press, 2010, ch. 16, pp. 397.

[21] Ethem Alpaydin, "Kernel Machines," in *Introduction to Machine Learning*, 2nd ed. Cambridge, MA: The MIT Press, 2010, ch. 13.

[22] Ethem Alpaydin, "Clustering," in *Introduction to Machine Learning*, 2nd ed. Cambridge, MA: The MIT Press, 2010, ch. 7, pp. 157.

[23] Stijn Van Dongen, "Graph Clustering via a Discrete Uncoupling Process," *Siam J. Matrix Anal Appl*, vol. 30, pp 121-141

[24] Chih-Wei Hsu et al., "A Practical Guide to Support Vector Classification," http://www.csie.ntu.edu.tw/~cjlin

[25] Bowman, A.W. and Azzalini, A., *Applied Smoothing Techniques for Data Analysis*, Clarendon Press, 1997, pp 50.

[26] Kapranov, P., "From transcription start site to cell biology," *Genome Biology*, vol. 10, pp 217-221.

[27] Tomasi, C., "Estimating Gaussian Mixture Densities with EM – A Tutorial," [online]. Available: http://www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf

VITA

Gregory Shannon Nichols, received B.S. in Physics from University of Missouri-Kansas City. He also received a Master of Business Administration from Bellevue University. He is pursuing M.S. in Computer Science at University of Missouri Kansas City. His research interests include machine learning/computational intelligence and applications in bioinformatics.