



Discovering the Distribution of Palindromic Sequences in the SMAD4 Gene using Large and Medium Deletions and the Resulting RNA Structure Predictions

Andrew Ninh^{1*} and Lisa Battig²

Student¹, Teacher²: Fountain Valley High School, 17816 Bushard Street, Fountain Valley, CA 92708

*Corresponding author: andrewninh@ieee.org

Abstract

The SMAD4 gene codes for cell-signaling proteins that prevent abnormal vascular growths. DNA palindromes are inversely proportional sequences that play roles in gene expression through the formation of stem-loops and disease/tumor detection. Previous research approximated that there were 100 palindromes in every 1000 base pairs of a randomly generated sequence. A Java program was written to mutate fasta sequences based on ARUP's SMAD4 database, information from NCBI, and random locations and another was written to find palindromes in the DNA sequences and output their lengths. These lengths were plotted sequentially using the Mathematica software. By measuring shifts in each mutated plot superimposed on the wild type plot, the number of pixels shifted between peaks was recorded and, using a scale which was measured to be 12 pixels per 100 palindromes, converted into the number of palindromes deleted. The amount of base pairs (bp) deleted was proportional to the amount of palindromes deleted. The relationship between bp and palindromes is described by the equation $p = \text{round}(-0.242996 + 0.425309 * l)$ such that p represents the number of palindromes and l represents the length of a sequence. This linear regression shows that palindromes are evenly distributed throughout the SMAD4 gene assuming the distribution follows a Poisson distribution. Out of every 1000 bp, there are approximately 420 palindromes in the SMAD4 mRNA which is approximately 475 palindromes in SMAD4's Primary Assembly genomic region. The SMAD4 gene exhibits 275 (137.5%) more palindromes than the randomly generated palindromic distribution projected by previous research. Finding the distribution of palindromes in RNA molecules can lead to future research and classification of key regions that determine the shape of secondary, tertiary, and quaternary structures.

Introduction

Palindromic sequences in DNA are sequences of nucleotides that are identical when read from the 5' to 3' end on one strand and 3' to 5' end of the complementary strand (and vice versa). Genetic palindromic sequences, whose latter half is the reverse complement of the former half, differ from lexical palindromes which read exactly the same forwards and backwards.

Palindromic sequences have a tendency of creating hairpin loops which are secondary RNA structures. These hairpin loops (also known as stem-loops) have many purposes which include protein synthesis repression¹. Nonrandom distribution patterns

of DNA palindromes have also been found in cancer cells and are attributed to mycogene amplification^{2,3}. Secondary, tertiary, and quaternary structures of RNA are important to its function and expression; if an RNA molecule was mutated by a deletion or other mutation (particularly large mutations), RNA expression will be compromised as it may be too susceptible to enzymes or it may be unable to unwind during expression.

Small mutations involve the deletion of a few base pairs and can cause frameshifts (unless a multiple of 3 base pairs are deleted). Large mutations can range from the deletion of entire genes to exons of base pairs and tend to be more detrimental to the functioning of a gene.

The SMAD4 gene (MADH4, Mothers against decapentaplegic homolog 4) codes for the SMAD4 protein. The SMAD4 protein suppresses tumors through cell signaling by regulating the TGF β proteins⁴. The protein also recognizes and binds to sites with the 8 bp palindromic sequences (GTCTAGAC). Mutations in this gene cause many cancers (50% of pancreatic cancers involve a mutated SMAD4 gene), particularly colorectal cancers⁵.

In addition to the gene's connection with cancers, SMAD4 mutations are also common in autosomal dominant syndromes such as Juvenile Polyposis Syndrome (JPS), Hereditary Hemorrhagic Telangiectasia (HHT), and a hybrid of the two diseases (JPS-HHT)^{6,7}. JPS is characterized by small benign polyps which can increase a patient's risk of gastrointestinal cancers^{8,9}. HHT is characterized by arteriovenous malformations (AVMs). AVMs are abnormal connections of veins and arteries that clump together. AVMs can occur in various parts of the body but usually occur in the nervous system. Patients with AVMs are at an increased risk of bleeding. This is especially fatal if the AVM is located in the brain or spinal cord.

The purpose of this study is to find a possible correlation and proportionality between the amount of deleted base pairs of a large deletion mutation and the resultant number of palindromes removed and stating this correlation in an equation. By finding how many palindromic sequences were deleted in each large mutation and comparing this ratio with palindromic deletions by medium deletions, an accurate number of palindromes in however many base pairs were found. The results only account for perfect palindromic sequences, not approximate palindromes. It is important to find DNA palindromes in RNA because of tendencies of forming hairpin loops and pseudoknots which are key components of secondary, tertiary, and quaternary structures; by finding concentrated spots of palindromes, further research can be done into classifying specific regions of RNA molecules that are key in RNA functioning.



By using large and medium deletion mutations an accurate rate at which palindromes are distributed throughout the gene were found. This is based upon the assumption that palindromes are uniformly distributed throughout the gene through a Poisson distribution¹⁰. If the amount of palindromes per base pair is proportional, then palindromes are distributed similarly to palindromic distributions in randomly generated and short genomes (such as that of bacteria)¹¹.

Deletions in the SMAD4 gene, which cause frameshift mutations, cause a shift (by deletion of palindromes) in the distribution of palindromic sequences that is proportional to the number of base pairs deleted. The shift in palindrome sequence distribution is similar to the frameshift in a nucleotide sequence following deletions and other frameshift mutations. As deletions occur and palindromic sequences are removed, the remaining palindromic sequences are shifted leftwards.

Materials and Methods

The appropriate .fasta files from GenBank were downloaded: Homo sapiens Chromosome 18, GRCh37.p5 Primary Assembly (region: 48,556,583 to 48,611,412) and Homo sapiens SMAD family member 4 (SMAD4)¹² (accession number NM_005359). The large deletion mutations on the SMAD4 database were converted into a readable .csv file (6 columns: (1) Segment (exon, etc.) (2) Type of mutation (3) Start location of the mutation (integer) (4) End location of the mutation (integer) (5) Nucleotide change (6) Type of protein change) titled "mutations.csv". Three Java programs were written: (Supplementary Data Index 1) reads mutation information from the mutations.csv file and accordingly mutates a .fasta file (Supplementary Data Index 2) finds palindromes and prints the length of each palindrome (length) (Supplementary Data Index 3) reads and writes .fasta files which are necessary file types for genetic sequences. Parameters were set on lengths of palindromes because 2 base pair (bp) palindromes are not important and a single nucleotide would not be a DNA palindrome. Due to the lack of large deletion mutations from the ARUP database, randomly chosen coding sequences (such as exons) were chosen to be deleted and added to the mutations.csv file. The information is divided up by columns and stored into 6 arrays (this way, the element [0] of each array would pertain to the first mutation and so on). The ListLinePlot function on Mathematica was used to plot the sequential lengths (sequence x length) of both the Primary Assembly and mRNA palindromes (see figures 1-2). Each respective plot's number of palindromes was counted. The SMAD4 mRNA .fasta file and the mutation-list on the .csv file were run through a Java program which created and saved 15 mutated .fasta files. The 15 mutated mRNA sequences were then run through the palindrome-finding program. Each sequence resulted in a string of numbers and commas (in the format "x, "). The outputs were saved into respective .csv files to be graphed. The .csv files are in a Cartesian system: the x-coordinates are the sequential appearance of each palindrome in the gene and the y-coordinates are the lengths of each palindrome. Lengths are chosen as the chosen y-coordinate because they are unchanging (unlike frequency) and would thereby make comparison and change measurement more accurate. Afterwards, a Mathematica program was written and employed the ListLinePlot function to plot the sequential lengths and then superimposed the mutated lengths graphs on the non-mutated lengths graphs (Supplementary Data Index Index 4). The lengths and frequencies superimpositions were saved in .png files. The .png image files were analyzed using the ImageJ program by using the straight-line tool to measure the "pixel shift" (the frame shift is measured by differences in pixels). The two superimposed graphs should be identical except one is shifted leftwards due to the deletion mutation. This process was repeated for different conspicuous peaks (multiple trials whose average value were recorded). Using the x-axis (increments of 10 occurrences) as a scale, pixel differences were measured which were used to compare and measure each pixel shift between peaks. By using accurate pixel-palindrome scales, the number of palindromes deleted per pixel shift was determined. These proportions were recorded and plotted on Mathematica to find a graph of best fit to determine a mathematical model of the distribution of palindromes.

Results

Figures 1-2 are plots of the number of palindromes in the SMAD4 mRNA and the Primary Assembly region. The Primary Assembly has 24,653 palindromes while mRNA has 3,738. The Primary Assembly was measured to contain on average, using random regions, 6.6 times as many palindromes as mRNA. It is important to note the amount of large palindromes found in the Primary Assembly compared to those found in the mRNA. Figures 3 and 4 show the overlap of wild type SMAD4 and deletions found in JPS and pancreatic cancer.

Tables 1-2 demonstrate that the A mutations are large mutations from the SMAD4 database¹³, B mutations are medium deletions generated from the list of coding sequences provided by GenBank, and C mutations are randomly generated deletions based on the location of large coding sequences/exons.

By measuring each plot using the ImageJ program, it was determined that 12 pixels represented 10 palindromes. This information was used as a scale to convert the distance between peaks in Table 1 from pixels to palindromes (see Table 2).

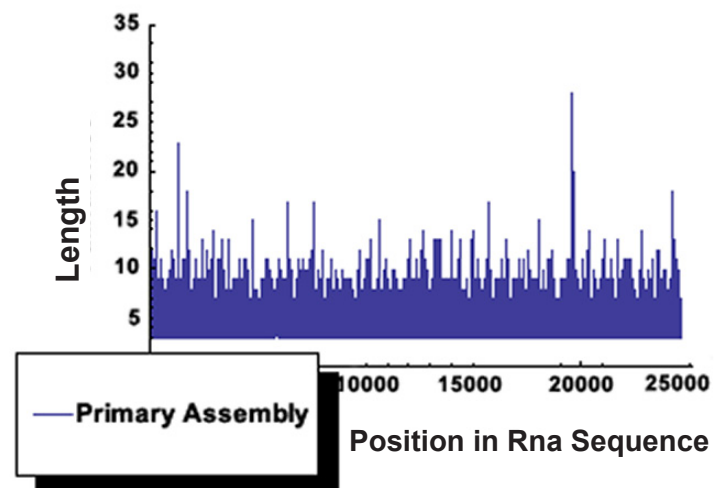


Figure 1. Palindromes in the SMAD4 Primary Assembly.

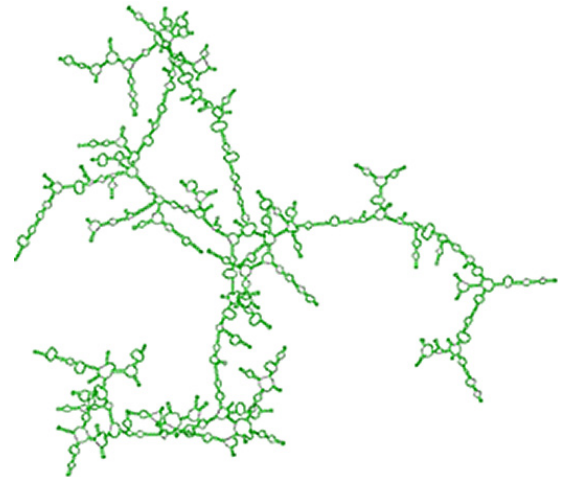
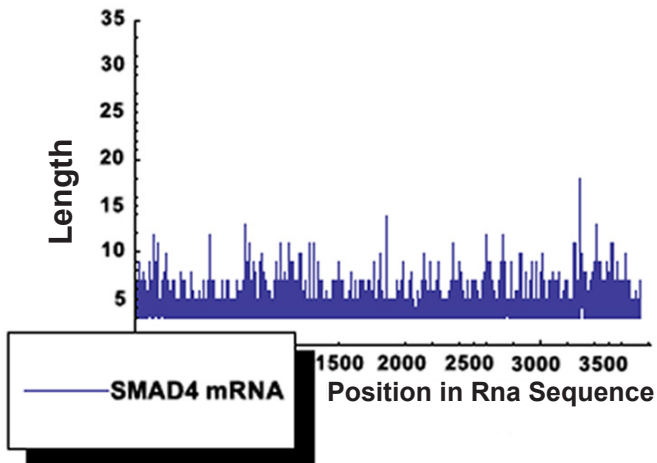


Figure 2. Palindromes in the SMAD4 mRNA and its predicted RNA structure (Free Energy of Structure = -1497.4 kkal/mol.).

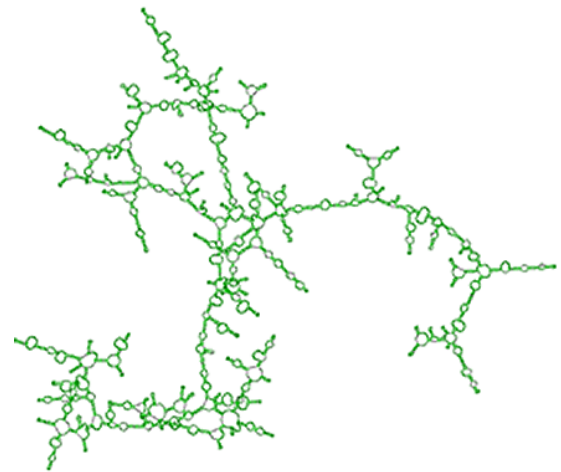
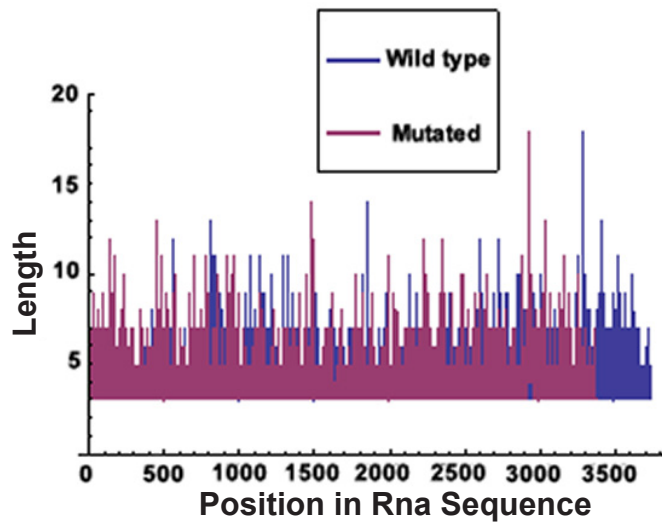


Figure 3. Palindromes in the SMAD4 mRNA 668 - 1659bp deletion mutant and its predicted RNA structure (Free Energy of Structure = -1309.0 kkal/mol.). This mutant has been shown to be involved in Juvenile Polyposis Syndrome.

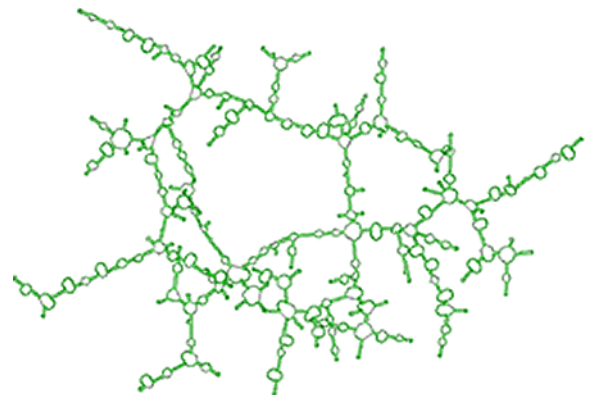
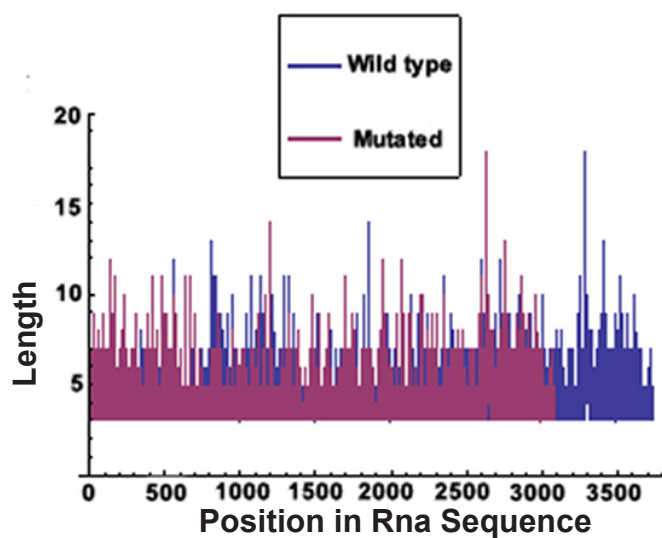


Figure 4. Palindromes in the SMAD4 mRNA 539 - 2197bp deletion mutant and its predicted RNA structure (Free Energy of Structure = -1181.3 kkal/mol.). This mutant has been shown to be involved in pancreatic cancer.



Table 1. SMAD4 deletions with peaks measured in pixels.

Location	100 seq	Peak 1	Peak 2	Peak 3	Peak 4	Peak 5	Average
A.							
1 - 1659 bp	12	87	87	87	87	87	87
668 - 1659 bp	12	44	44	44	44	44	44
788 - 1659 bp	12	39	39	39	39	40	39
B.							
1-411 bp	12	30	30	30	30	30	30
993-1205 bp	12	8	8	8	8	8	8
1326-1442 bp	12	5	5	5	5	5	5
1678-1985 bp	12	17	17	17	-	-	17
2277-3086 bp	12	45	46	45	45	45	45
5401-5563 bp	12	8	7	8	8	8	8
7506-7796 bp	12	17	17	17	18	17	17
C.							
539 - 2197 bp	12	79	79	79	79	79	79
2197 - 3865 bp	12	89	89	89	-	-	89
3865 - 5524 bp	12	85	85	85	85	-	85
5524 - 7183 bp	12	80	80	80	80	-	80
1986-8772 bp	12	347	-	-	-	-	347

Table 2. SMAD4 deletions with peaks covered from pixels to palindromes.

Location	Length	Peak 1	Peak 2	Peak 3	Peak 4	Peak 5	Average
A.							
0 - 1659 bp	1659	730	730	730	730	730	730
668 - 1659 bp	991	370	370	370	370	370	370
788 - 1659 bp	871	330	330	330	330	330	330
B.							
1-411 bp	411	250	250	250	250	250	250
993-1205 bp	212	70	70	70	70	70	70
1326-1442 bp	116	40	40	40	40	40	40
1678-1985 bp	397	140	140	140	-	-	140
2277-3086 bp	809	380	380	380	380	380	380
5401-5563 bp	162	70	60	70	70	70	70
7506-7796 bp	290	140	140	140	150	140	140
C.							
539 - 2197 bp	1658	660	660	660	660	660	660
2197 - 3865 bp	1668	740	740	740	-	-	740
3865 - 5524 bp	1659	710	710	710	710	-	710
5524 - 7183 bp	1659	670	670	670	670	-	670
1986-8772 bp	6786	2890	-	-	-	-	2890

Table 2 shows the same exact data as Table 1 but the information is converted from pixels to palindromes. Length and Average deleted palindromes were plotted and the line of best fit would determine a mathematical model of the number of palindromes in a sequence of base pairs.

Each length of each deletion in bp by the resultant average number of palindromes deleted were plotted (see Figure 5) and the line of best fit, $p = \text{round}(-0.242996 + 0.425309 * l)$, which follows a linear regression, was determined.

In addition, to examine the changes in the RNA secondary structures, the wild type and mutated fasta files were analyzed in GeneBee Molecular Biology Server's RNA secondary structure prediction program and the resulting structures and stem energy were recorded adjacent to each respective plot.

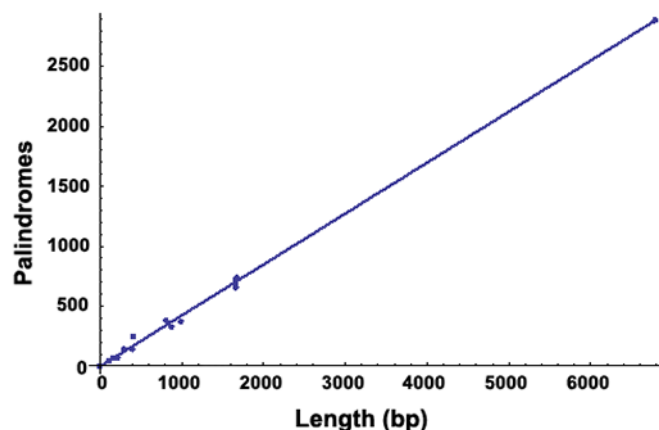


Figure 5. Number of palindromes by sequence length.

Discussion

The linear regression model found that, in mRNA, the SMAD4 gene exhibited approximately 4200 palindromes per 1000 bp. Given this, the Primary Assembly has about 1.14 times as many mRNA palindromes (and assuming palindromes are fairly evenly distributed); about 475 palindromes were calculated to exist in every 1000 bp. This model was found to be accurate within a range of about 13 palindromes (~5%).

The shift in pixels between graphs was the desired measurement rather than simply counting palindromes of each mutated sequence and subtracting that number from the wild type number of palindromes (3,738) because one trial might not provide sufficient information and there might be new palindromes formed as a result of a deletion.

According to the SMAD4 database, the three large mutations ended at bp 1659. This might be caused by unique tendencies of sequences in that exon/CDS to form cruciform or hairpin loops and become cut from the strand. The majority of deletion mutations (10 out of 15) were randomly chosen exons/coding sequences. These sequences might possess a certain quality that renders them unviable deletions.

Technically, the methods are accurate though pixel recording has to be as accurate as possible. If a pixel is only slightly off for the pixel-palindrome scale, entire results could end up being several palindromes inaccurate. Errors and variation in measuring pixel shifts are usually insignificant. However, large errors and inconsistencies demonstrate changes in palindrome distribution.

The distribution of palindromic sequences is best described as a linear regression model which assumes that there is no correlation between palindromic occurrences and there is a constant variance (Poisson distribution). The linear equation: $p = \text{round}(-0.242996 + 0.425309 * l)$ such that p represents the number of palindromes and l is the length of a sequence satisfies this linear regression model.

The equation counts palindromes indiscriminately (counting palindromes within palindromes) and does not measure uniformity nor proximity of palindromes. The results showed that the approximate number of palindromes stated in previous research¹¹ – 200 palindromes every 1000 randomly generated bp – was inaccurate. According to the equation, the average number of palindromes every 1000 bp is actually 475 (137.5% more).

The mathematical model was effectively equated by measuring and plotting the number of palindromes that were shifted due to a large or medium deletion. Only a few deletions were referenced from the ARUP database; the randomly deleted coding sequences may not be fully accurate or viable.



Ultimately, this study provides the first mathematical model of palindromic sequence distribution in the SMAD4 gene and found an accurate proportion of 475 palindromes (Primary Assembly) or 435 palindromes (mRNA) for every 1000 base pairs. By finding an accurate depiction of how palindromes are distributed throughout the SMAD4 gene, further research into palindromic sequences is possible. By knowing how many palindromes exist in however many base pairs, one can find patterns in uniformity of palindromes (or specific palindromes).

The Java program used could be revised to include approximate palindromes. In the future, distribution patterns between approximate and perfect palindromes may be compared to determine whether both approximate and perfect palindromes follow a linear regression or if they both follow a similar equation. A universal mathematical model for palindromic distribution has not yet been found. In addition, the mathematical model might be used in the future to compare and contrast numbers of palindromes of different genes, genomes, and randomly generated sequences. For example, by comparing the variance of various genes and genomes (which require certain palindromes to function) with randomly generated sequences (which should be purely random), evolutionary patterns in genes can be traced.

References

1. Drew, H. R., Lewy, D., Conaty, J., Rand, K. N., Hendry, P. H., & Lockett, T. (1999). RNA hairpin loops repress protein synthesis more strongly than hammerhead ribozymes. *The European Journal of Biochemistry* 266(1): 260-73.
2. Choi, C. Q. (2005). DNA palindromes found in cancer. *Genome Biology* 6(2).
3. Tanaka, H., Bergstrom, D., Yao, M., & Tapscott, S. (2005, February 13). Widespread and nonrandom distribution of DNA palindromes in cancer cells provides a structural platform for subsequent gene amplification. *Nature Genetics* 37(3): 320-7.
4. Zawel, L., Dai, J., Buckhaults, P., Kinzler, K., Vogelstein, B., & Kern, S. (1998, February 17). Human Smad3 and Smad4 are sequence-specific transcription activators. *Molecular Cell* 1(4): 611-7.
5. Woodford-Richens, K., Rowan, A., Gorman, P., Halford, S., Bicknell, D., Wasan, H., et al. (2001, June 25). SMAD4 mutations in colorectal cancer probably occur before chromosomal instability, but after divergence of the microsatellite instability pathway. *Proceedings of the National Academy of Sciences of the United States of America* 98(17): 9719-23.
6. Gallione, C., Aylsworth, A., Faughnan, M., Ganguly, A., Garvie, J., Henderson, K., et al. (2010). Overlapping spectra of SMAD4 mutations in juvenile polyposis (JP) and JP-HHT syndrome. *American Journal of Medical Genetics* 152A(2): 333-9.
7. Iyer, N. K., Burke, C. A., Leach, B. H., & Parambit, J. G. (2010). SMAD4 mutation and the combined syndrome of juvenile polyposis syndrome and hereditary haemorrhagic telangiectasia. *Thorax* 65 (8): 745-6.
8. Moguelet, P., Plassa, L., Metayer, J., Teniere, P., Olschwang, S., Frebourg, T., et al. (2003). SMAD4-related Familial Juvenile Polyposis Syndrome with Colon Cancer. *Cancer Cancer Genomics and Proteomics* 1(1): 33-38.
9. Schwenter, F., Faughnan, M., Grandinger, A., Berk, T., Gryfe, R., Pollett, A., et al. (2012). Juvenile polyposis, hereditary hemorrhagic telangiectasia, and early onset colorectal cancer in patients with SMAD4 mutation. *Journal of Gastroenterology* 47(7): 795-804.
10. Leung, M., Choi, K. P., Xia, A., & Chen, L. H. (2005). Nonrandom Clusters of Palindromes in Herpesvirus Genomes. *Journal of Computational Biology* 12(3): 331-54.
11. Goh, Y. K. (n.d.). Distribution of DNA Palindromes in Short Palindromes. *Symposium on Progress in Information & Communication Technology*.
12. Wooderchak, W. L., Spencer, Z., Crockett, D. K., McDonald, J., & Bayrak-Toydemir, P. (2010). Repository of SMAD4 Mutations: Reference for Genotype/ Phenotype Correlation. OMICS Group:Open Access Publisher & Scientific Events Organizer 1(1): 101.
13. Aretz, S., Stienen, D., Stolte, M., Entius, M., Loff, S., Back, W., et al. (2007). High proportion of large genomic deletions and a genotype-phenotype update in 80 unrelated families with juvenile polyposis syndrome. *Journal of Medical Genetics* 44 (11): 702-9.

Acknowledgements

I would like to thank and acknowledge the University of Utah's ARUP Scientific Resource for Research and Education for providing the open-source SMAD4 database. I would like to also thank and acknowledge GeneBee Molecular Biology Server's RNA secondary structure prediction program for providing me with the stem energy and image of the RNA secondary structure.



Supplementary Data

Index 1

```
import java.util.*;
import java.io.*;
public class MutateGene { //arrays can be initialized in the
constructor
    private String [] segment = new String [],
        mutationType = new String [],
        nucMutation = new String [],
        proteinChange = new String [];
    private int [] changeLoc1 = new int [],
        changeLoc2 = new int [];

    /**
     * Constructor
     */
    public MutateGene(String fileName) throws IOException {
        BufferedReader b = new BufferedReader(new FileReader
(fileName));
        Scanner input = new Scanner(b);
        String info = "";
        while(input.hasNextLine())
            info += input.nextLine().trim() + ";";
        StringTokenizer tokenizer = new StringTokenizer(info, ";");
        String cell [] = new String [/*number of desired columns
* 6 */];
        for(int i = 0; tokenizer.hasMoreTokens(); i++)
            cell[i] = tokenizer.nextToken().trim();
        for(int i = 0; i < cell.length; i++) {
            if((i%6)==0)
                segment[i/6] = cell[i];
            else if((i%6)==1)
                mutationType[i/6] = cell[i];
            else if((i%6)==2)
                changeLoc1[i/6] = Integer.parseInt(cell[i]);
            else if((i%6)==3)
                changeLoc2[i/6] = Integer.parseInt(cell[i]);
            else if((i%6)==4)
                nucMutation[i/6] = cell[i];
            else
                proteinChange[i/6] = cell[i];
        }
    }
    void mutation(String DNA) throws IOException {
        Mutate m = new Mutate();
        for(int i = 0; i < /*number of columns*/; i++) {
            if(mutationType[i].equals("Deletion"))
                m.deletion(DNA, segment[i], nucMutation[i],
                    changeLoc1[i], changeLoc2[i],
proteinChange[i]);
                //Other unused commands
        }
    }
}
import java.io.*;
public class Mutate {
    private FASTAFile newFile = new FASTAFile();
```

```
private String name, description, sequence;
void deletion(String DNA, String segment, String nucleotide,
int loc1,
int loc2, String protein) throws IOException {
    sequence = DNA.substring(0,loc1) +DNA.
substring(loc2+1,DNA.length());
    name = "Deletion_at" + loc1 + "_" + loc2 + ".fasta";
    description = ">gi|(deletion) Homo sapiens chromosome
18, GRCh37.p5 "
+ "changes at: " + segment + "(" + loc1 + "-" + loc2 + ")" +
"protein change: " + protein;
    newFile.writeFASTA(name, description, sequence, seg-
ment);
}
//Other unused methods
}
```

Index 2

```
import java.io.*;
import java.util.*;
public class FASTAFile {
    public String readFASTA(String fname)throws IOException {
        BufferedReader r = new BufferedReader(new
FileReader(fname));
        Scanner input = new Scanner(r);
        String dna = "";
        String sDescription = input.nextLine();
        while(input.hasNext()) {
            dna += input.nextLine().trim();
        }
        return dna;
    }
    public void writeFASTA(String fName, String sDescription,
String
sSequence, String seqLoc)throws IOException {
        String fLoc = //desired location
        PrintWriter out = new PrintWriter(new BufferedWriter(new
FileWriter(fLoc + fName)));
        int nStart=0;
        int nNum = 70;
        out.println(sDescription);
        while (nStart < sSequence.length()) {
            if (sSequence.length() - nStart < nNum)
                out.println(sSequence.substring(nStart));
            else
                out.println(sSequence.substring(nStart, nStart+70));
            nStart = nStart + nNum;
        }
        out.close();
    }
}
```

Index 3

```
import java.util.*;
public class DnaPalindromes {
    // Finds palindromic DNA sequences and counts them.
    // The variable i is the first boundary of the array and starts at
```



```

0 and
// ends 1 before the last nucleotide because
// a single nucleotide is not considered a palindrome.
void findPalindromes(String dna){
    for(int i = 0; i < dna.length() - 1; i++){
        for (int j = i + 1; j < dna.length(); j++){
            String palSeq = dna.substring(i,j+1);
            int count = 0;
            if(isPalindromes(i,j,dna) && palSeq.length()>=3)
                //Necessary command
        }
    }
}
boolean isPalindromes(int i, int j, String palArr){
    DnaCompare dnaComp = new DnaCompare();
    if(i > j - 1)
        return true;
    if(!dnaComp.correctPair(palArr.charAt(i), palArr.charAt(j)))
        return false;
    return isPalindromes(i+1, j-1, palArr);
}
public class DnaCompare {
    // Finds the correct base pairings only for exact palindromes
    boolean correctPair(char base1, char base2){
        if((base1 == 'A' && base2 == 'T') ||
            (base1 == 'T' && base2 == 'A') ||
            (base1 == 'C' && base2 == 'G') ||
            (base1 == 'G' && base2 == 'C')){
            return true;
        }
        else
            return false;
    }
}

```

Index 4

```
SetDirectory["C:\\Users\\andrew\\Documents\\Bioinformatics\\SMAD4\\SMAD4_grapher"]
```

```
Show[ListLinePlot[Take[Import["name.csv"], 1],PlotRange->{0,20},PlotStyle->Color],
ListLinePlot[Take[Import["name.csv"], 1],PlotRange->{0,20},PlotStyle->Color]]
```