Six-Letter Words in DNA

By Craig Paardekooper

Data Sources

Pan troglodytes genome assembly NHGRI_mPanTro3-v2.1_pri - NCBI - NLM

Homo sapiens genome assembly T2T-CHM13v2.0 - NCBI - NLM

The counts I have carried out previously were for codons (three letter words), so extending this to 4 letter words or more is a natural progression.

Rather than counting codons, which are words made of 3 nucleotide letters, I decided to count the frequencies of every word of length 6 nucleotides. There are 4096 different words made of 6 letters, or 4^6.

Usually only select sequences are compared – which leads to biased results. Here I compare the entire chromosomes.

Method

First, I created and populated an array with all 4096 possible 6 letter words. Then I looped through the DNA of the human Y chromosome in steps of 6 letters, and incremented the array by 1 each time a particular word occurred. I did the same for the chimp Y chromosome. I then copied and pasted the results into excel and counted the differences between the counts for chimpanzee and human DNA.

Results

6-letter word frequencies in the Y chromosome can be viewed here – https://howbad.info/6-letter-words.xlsx

7-letter word frequencies in the Y chromosome can be viewed here – https://howbad.info/7-letter-words.xlsx

7-letter word frequencies in the X chromosome can be viewed here – https://howbad.info/7-letter-words-X.xlsx

Observations for 6-letter-word frequencies in Y Chromosome

There are 10.4 million 6-letter words in the human Y chromosome
There are 6.07 million 6-letter words in the chimp Y chromosome
The human Y chromosome is 71% bigger than the chimp Y chromosome

Stats for the Human Y Chromosome
1. 700 (17%) of the 6-letter words, occur with more than double the frequency compared to in the chimp Y chromosome
These 700 words make up more than half of the entire human chromosome Y
These 700 words make up only one sixth of the entire chimp chromosome Y

2. 300 (7.3%) of the 6-letter words, occur with more than triple the frequency compared to in the chimp Y chromosome
These 300 words make up 36% of the entire human Y chromosome
These 300 words make up only 6% of the entire chimp Y chromosome

4. 230 (5.6%) of the 6-letter words, occur with more than 4 times the frequency compared to in the chimp Y chromosome
These 230 words make up more than one third of the entire human Y chromosome
These 230 words make up only 4.7% of the entire chimp Y chromosome

Observations for 7-letter word frequencies in the Y Chromosome

For 7 letter words the differences between the human and chimp chromosome Y are even more extreme

2345 words out of 16384 (14.31 %) occur with more than double the frequency compared to in the Chimp Y chromosome -

- These words make up 5312597 of the 8919099 7-letter-words in the human Y chromosome - that's 60% of the Y chromosome
- These words make up 877738 of the 5205300 7-letter-words in the chimp Y chromosome - that's 16% of the Y chromosome

So, the 7 letter words that make up 60% of the human Y chromosome, only make up 16% of the chimp Y. This indicates that we are not 98% identical to chimpanzees.

Observations for 7-letter word frequencies in the X Chromosome

There are 22.02 million 7-letter words in the human X chromosome
There are 21.98 million 7-letter words in the chimp X chromosome

The two chromosomes are therefore almost identical size

Despite the larger size compared to the Y chromosome –

- only 63 words in human X (0.38% of 16384) occur with double the frequency compared to chimp X
- these 63 words make up only 1/1000 th of human X
- these 63 words make up only 1/2000 th of the chimp X

Compare this to the Y chromosome where -

- 700 words occur with double the frequency in human Y compared to chimp Y
- these 700 words make up 50% of the human Y
- these 700 words make up only 16% of the chimp Y

So, the difference between human Y and chimp Y is 500 times greater than the difference between human X and chimp X.

Code for 6-letter word frequencies

```
Public Class Form1

          Dim Count As Integer
          Dim N As Integer = 0
          Dim x As Integer
          Dim Multiline As String = ""

Private Sub Button2_Click(sender As Object, e As EventArgs) Handles Button2.Click

        Count = 0
        Dim path As String = "C:\Users\craig\Downloads\Chromosomes\trogY.fasta"
        Dim Chromosome As String = "Chromosome2C"
        N = 600
        Dim sr As StreamReader = New StreamReader(path)
        Do While (sr.Peek() >= 0)
                Count += 1
                If Count Mod N <> 0 Then
                        Application.DoEvents()
                        Multiline &= sr.ReadLine
               Else
                        Multiline = Multiline.Replace(vbCrLf, "")
                        Multiline = Multiline.Replace(vbCrLf, "")
                        Multiline = Multiline.Replace(vbCrLf, "")
                        Multiline = Multiline.Replace(vbLf, "")
                        Multiline = Multiline.Replace(" ", "")
                        ProcessLines3(Multiline, Chromosome)
                        Multiline = ""
                End If
         Loop
         Dim results As String = ""
         For i = 0 To 4095
                results &= mArray(i) & vbTab & narray(i) & vbCrLf
         Next
         RichTextBox2.Text = results

End Sub
```

```
Sub ProcessLines3(MultiLine)
    Dim Bin As String = ""
    If MultiLine.Length > 6 Then
        For y As Integer = 0 To MultiLine.length - 6 Step 6
            Bin = MultiLine.Substring(y, 6)
            For i = 0 To 4095
                If mArray(i) = Bin Then
                    narray(i) += 1
                    nucleotides += 6
                    exit for
                End If
            Next
        Next
   TextBox1.Text = nucleotides
   End If
End Sub
    Dim narray(4095) As Integer
    Dim mArray(4095) As String
    Dim nucleotides As Long = 0
Sub Permute()
    Dim word As String = ""
    Dim numb As Integer = 0
    Dim array() As String = {"T", "C", "A", "G"}
        For Each l As String In array
            For Each l2 As String In array
               For Each l3 As String In array
                    For Each l4 As String In array
                        For Each l5 As String In array
                            For Each l6 As String In array
                                mArray(numb) = l & l2 & l3 & l4 & l5 & l6
                                numb += 1
                            Next
                        Next
                    Next
                Next
            Next
        Next
 End Sub
```

```
Private Sub Form1_Load(sender As Object, e As EventArgs) Handles MyBase.Load

        Permute()

  End Sub

End Class
```

Update of Code for Sub Processlines

This update makes the code 4 times faster

```
Sub ProcessLines4(Multiline, Chromosome)
    Dim Bin As String = ""
    Dim TwoLetter As String = ""
    If Multiline.Length > 6 Then
        For y As Integer = 1 To Multiline.length - 7 Step 7
            Bin = Multiline.Substring(y, 7)
            TwoLetter = Multiline.Substring(y, 2)

            Select Case TwoLetter
                Case "TT"
                    For i = 0 To 1023
                        If mArray(i) = Bin Then
                            narray(i) += 1
                            nucleotides += 7
                            Exit For
                        End If
                    Next
                Case "TC"
                    For i = 1024 To 2047
                        If mArray(i) = Bin Then
                            narray(i) += 1
                            nucleotides += 7
                            Exit For
                        End If
                    Next

                Case "TA"
                    For i = 2048 To 3071
                        If mArray(i) = Bin Then
                            narray(i) += 1
                            nucleotides += 7
                            Exit For
                        End If
                    Next
```

```
Case "TG"

        For i = 3072 To 4095

                If mArray(i) = Bin Then

                        narray(i) += 1

                        nucleotides += 7

                        Exit For

                End If

        Next


Case "CT"

        For i = 4096 To 5119

                If mArray(i) = Bin Then

                        narray(i) += 1

                        nucleotides += 7

                        Exit For

                End If

        Next

Case "CC"

        For i = 5120 To 6143

                If mArray(i) = Bin Then

                        narray(i) += 1

                        nucleotides += 7

                        Exit For

                End If

        Next


Case "CA"

        For i = 6144 To 7167

                If mArray(i) = Bin Then

                        narray(i) += 1

                        nucleotides += 7

                        Exit For

                End If

        Next
```

```
Case "CG"
    For i = 7168 To 8191
        If mArray(i) = Bin Then
            narray(i) += 1
            nucleotides += 7
            Exit For
        End If
    Next

Case "AT"
    For i = 8192 To 9215
        If mArray(i) = Bin Then
            narray(i) += 1
            nucleotides += 7
            Exit For
        End If
    Next
Case "AC"
    For i = 9216 To 10239
        If mArray(i) = Bin Then
            narray(i) += 1
            nucleotides += 7
            Exit For
        End If
    Next

Case "AA"
    For i = 10240 To 11263
        If mArray(i) = Bin Then
            narray(i) += 1
            nucleotides += 7
            Exit For
        End If
    Next
```

```
Case "AG"
    For i = 11264 To 12287
        If mArray(i) = Bin Then
            narray(i) += 1
            nucleotides += 7
            Exit For
        End If
    Next

Case "GT"
    For i = 12288 To 13311
        If mArray(i) = Bin Then
            narray(i) += 1
            nucleotides += 7
            Exit For
        End If
    Next
Case "GC"
    For i = 13312 To 14335
        If mArray(i) = Bin Then
            narray(i) += 1
            nucleotides += 7
            Exit For
        End If
    Next

Case "GA"
    For i = 14336 To 15359
        If mArray(i) = Bin Then
            narray(i) += 1
            nucleotides += 7
            Exit For
        End If
    Next
```

```
                Case "GG"
                    For i = 15360 To 16383
                        If mArray(i) = Bin Then
                            narray(i) += 1
                            nucleotides += 7
                            Exit For
                        End If
                    Next

            End Select
        Next
        TextBox1.Text = nucleotides
    End If
End Sub
```

I can further increase the speed by sub-dividing each case using third and fourth letters, but this wont be necessary unless I want to look at even longer letter sequences.

Contact

Craig Paardekooper

craig@howbad.info