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

Observations for 6-letter-word frequencies

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

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.

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
```

Contact

Craig Paardekooper

craig@howbad.info