**BS32011: Applied Bioinformatics Assignment 3: Sequence Alignment Analysis**

**Fiona Macfarlane (110010712)**

Part A. You should write some code in Python or R to parse these files (or any file of this type) so that you can plot different columns against each other: (4%)

I used a parser to create a data structure for each of the files, serprot\_pairs.out and ig\_pairs.out. This data structure is the table from the output file. The python scripts used , ig\_parse.py and serprot\_parse.py can be found in the following github repository <https://github.com/fmacfarlane/Bioinformatics/tree/master/assignment_3>.

Two text files, ig\_pairs\_parsed.txt and serprot\_pairs\_parsed.txt, contains the parsed data.

These data structures can then be imported into RStudio and the columns plotted against each other.

B. Having written your parser, explore the data you have read in. Some things to look at are:

Q1. What is the relationship between %IDENT, NAS and SCORE and how do they correlate?

Using RStudio the values for %IDENT, NAS and SCORE were plotted against each other. Where %IDENT is the percentage of the sequences that are similar in the alignment, if this value is between 0-20% it is not statistically relevant and could have occurred by chance rather than genetic relationships. The NAS value is the normalised alignment score, and the higher the score the stronger the alignment. NAS is equal to the MATCH score divided by the number of aligned positions multiplied by 100. SCORE is equal to (RMEAN-MATCH)/STDEV.

Ig Pairs Data

The following code was used in R to plot the columns against each other:

>IDENT <- ig\_pairs\_parsed[1:1188110,"X.IDENT"]

>NAS <- ig\_pairs\_parsed[1:1188110,"NAS"]

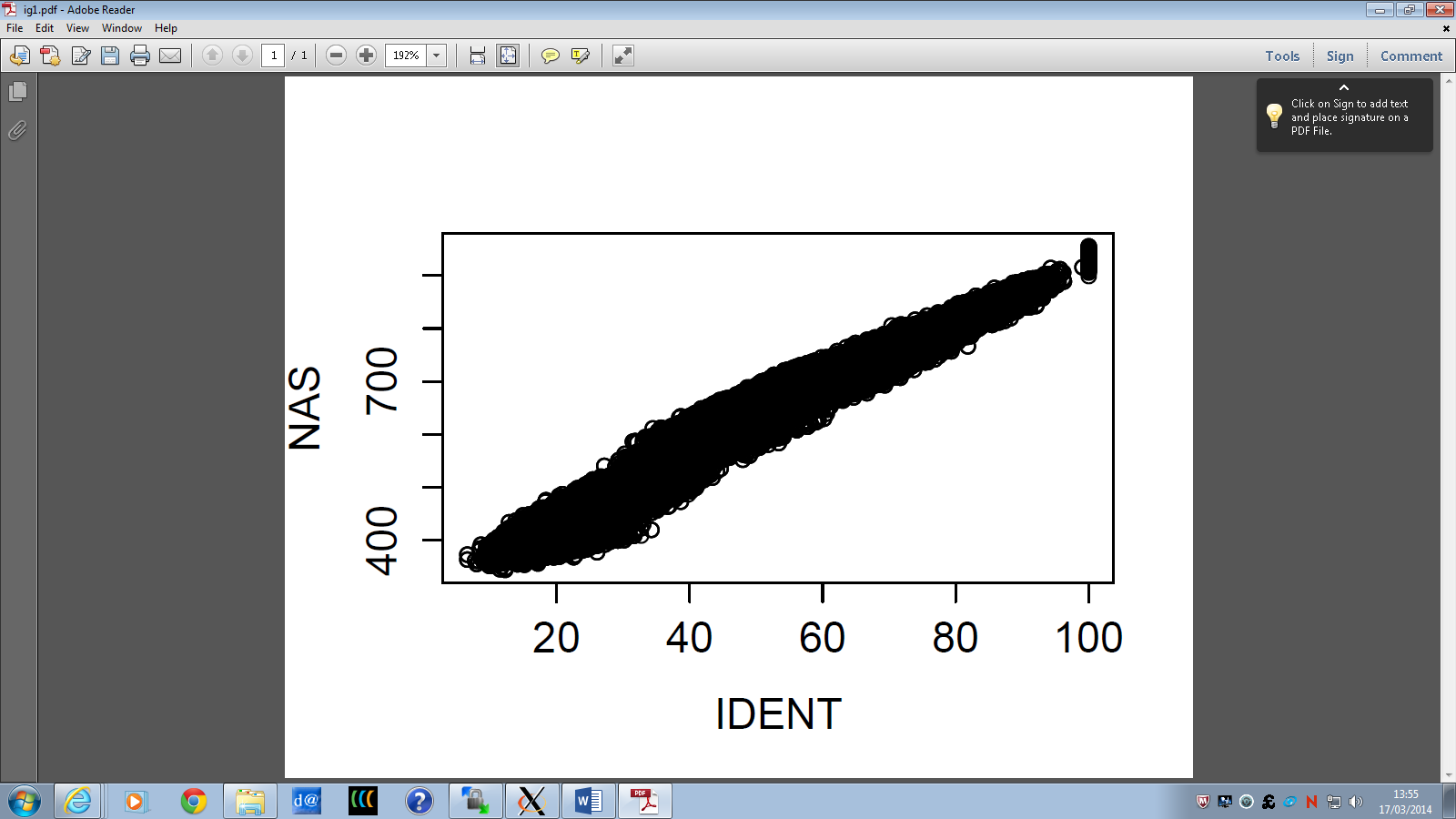
>SCORE <- ig\_pairs\_parsed[1:1188110,"SCORE"]

>plot(IDENT,NAS)#plot1i

>plot(IDENT,SCORE)#plot2i

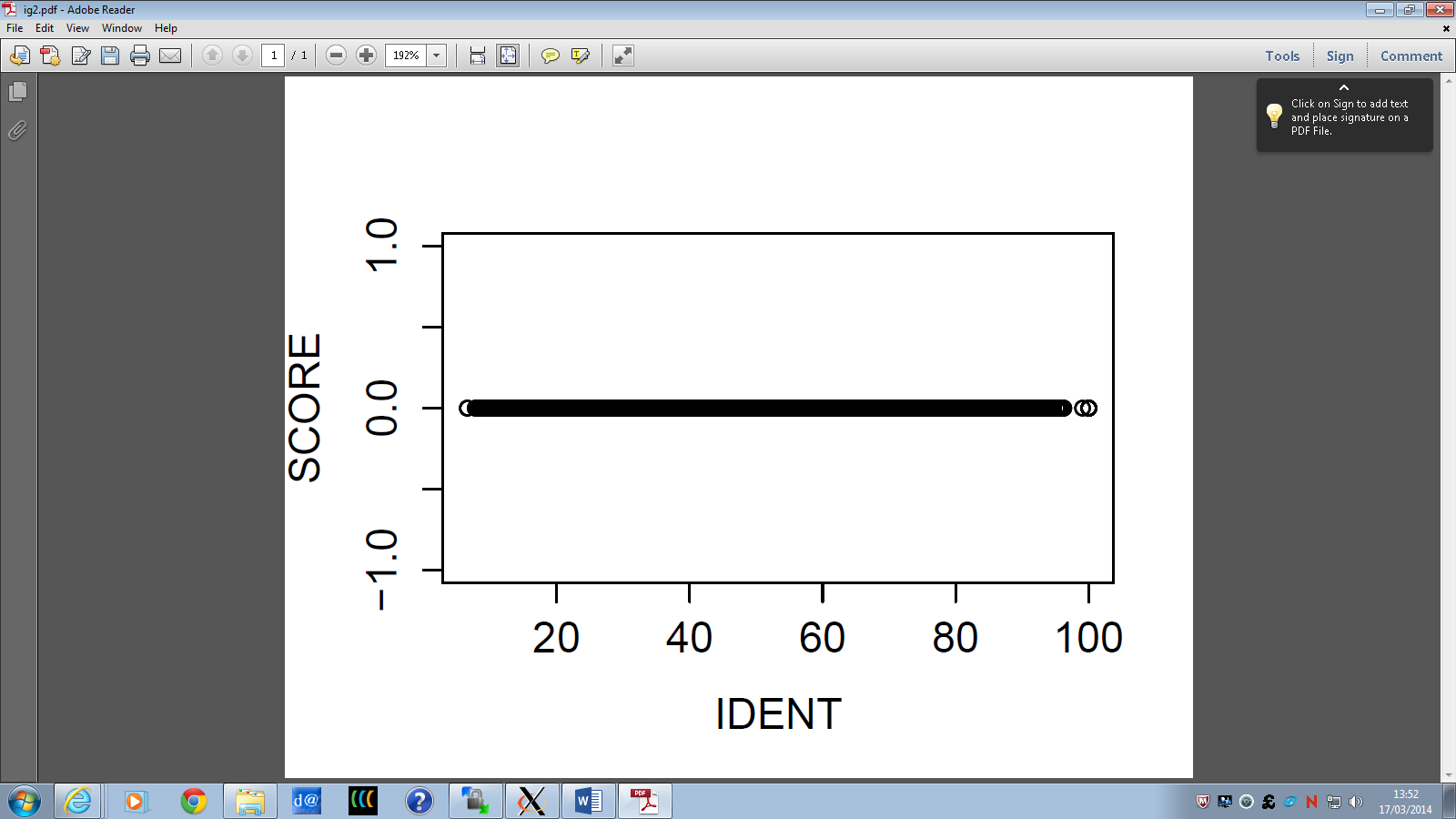
>plot(NAS,SCORE)#plot3i

Plot1i= IDENT vs NAS



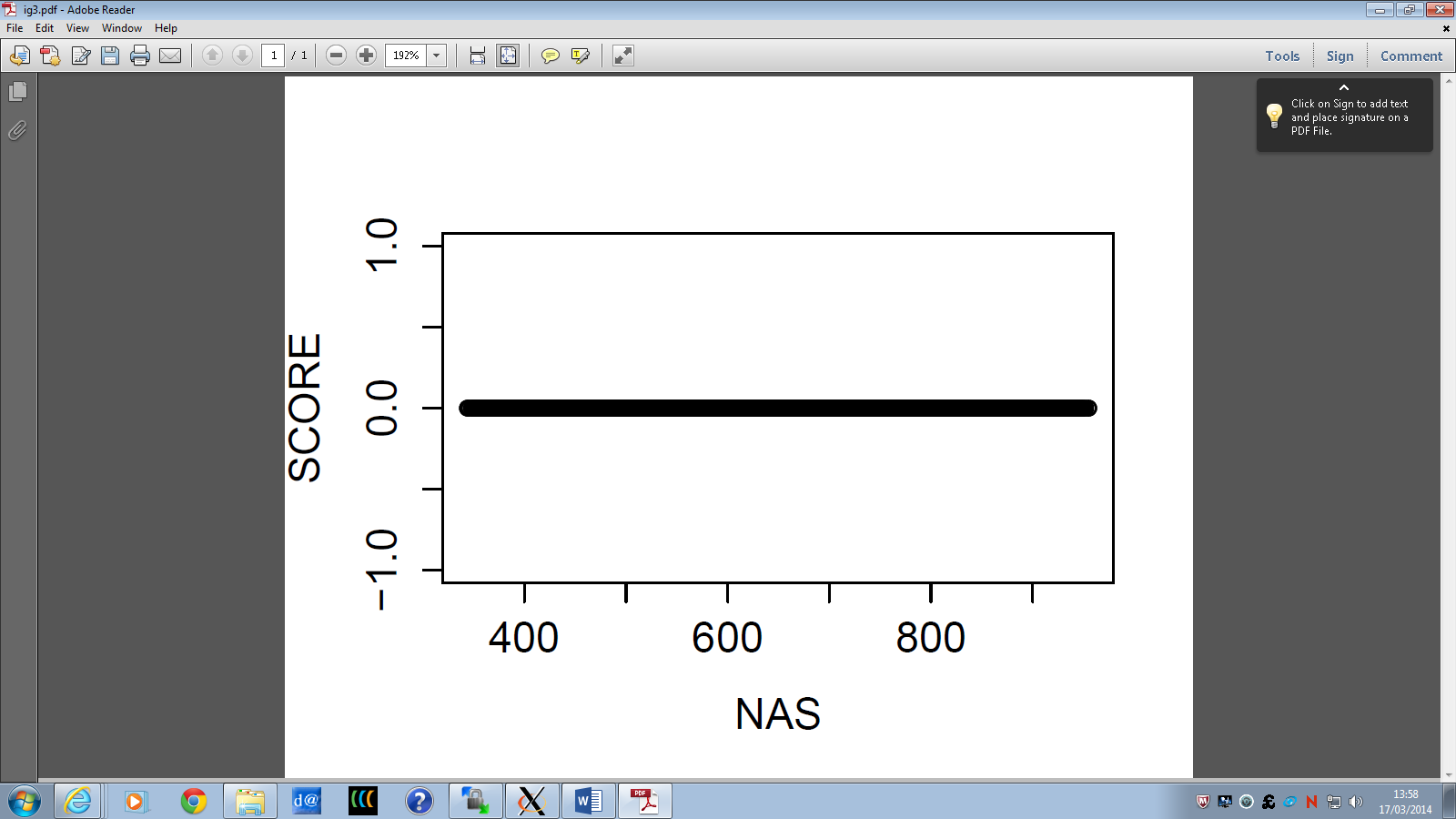
There is positive correlation between NAS and IDENT values. This suggests that the normalised alignment score relies on the percentage identity.

Plot2i= IDENT vs SCORE



The z-score of this file is always zero, suggesting that there is no randomisation taking place. Therefore the score will remain the same for increasing IDENT values.

Plot3i= NAS vs SCORE



The z-score of this file is always zero, suggesting that there is no randomisation taking place. Therefore the score will remain the same for increasing NAS values.

Serprot Pairs Data

This was done using the following code:

>serprot\_pairs\_parsed <- read.delim("~/Bioinformatics/Bioinformatics/assignment\_3/serprot\_pairs\_parsed.txt")

>View(serprot\_pairs\_parsed)

>x <- serprot\_pairs\_parsed[,"X.IDENT"]

>y <- serprot\_pairs\_parsed[,"NAS"]

>z <- serprot\_pairs\_parsed[,"SCORE"]

>plot(x,y)#plot1s

>plot(x,z)#plot2s

>plot(y,z)#plot3s

Plot1s: %IDENT(x) vs NAS(y)



From Plot1 it can be argued that generally as %IDENT increases, NAS increases. This makes sense since as the NAS score takes into account the %IDENT value.

PLot2s= %IDENT vs SCORE



Similarly to Plot1, Plot 2 shows that, generally, as %IDENT increases the SCORE increases. This again suggests that the higher the percentage identity the better the alignment.

Plot 3s=score vs NAS



Plot 3 again shows a positive trend between NAS and SCORE. The higher the NAS value the stronger the alignment. However a high NAS score does not suggest phylogenetic significance.

Q2. What is the relationship between sequence length (ILEN, JLEN) and MATCH, %IDENT and SCORE

Using RStudio the values for %IDENT, MATCH and SCORE were plotted against the ILEN values. Where %IDENT and SCORE are as above. The ILEN is the sequence length and the MATCH is the alignment score.

Ig pairs

The following code was used in R to plot the columns against each other:

>IDENT <- ig\_pairs\_parsed[1:1188110,"X.IDENT"]

>MATCH <- ig\_pairs\_parsed[1:1188110,"MATCH"]

>SCORE <- ig\_pairs\_parsed[1:1188110,"SCORE"]

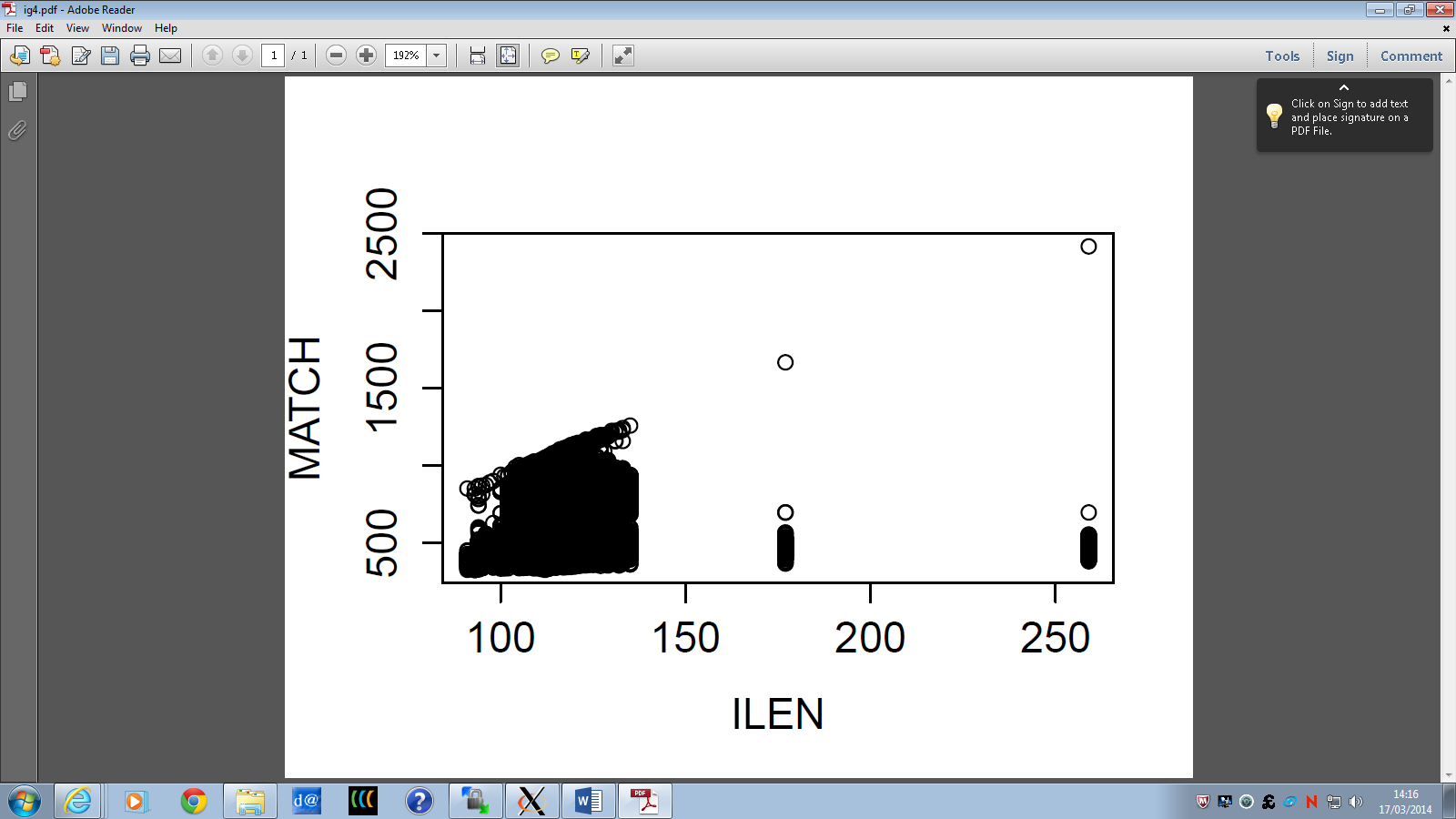
>ILEN <- ig\_pairs\_parsed[1:1188110,“ILEN”]

>plot(ILEN,MATCH)#plot4i

>plot(ILEN,IDENT)#plot5i

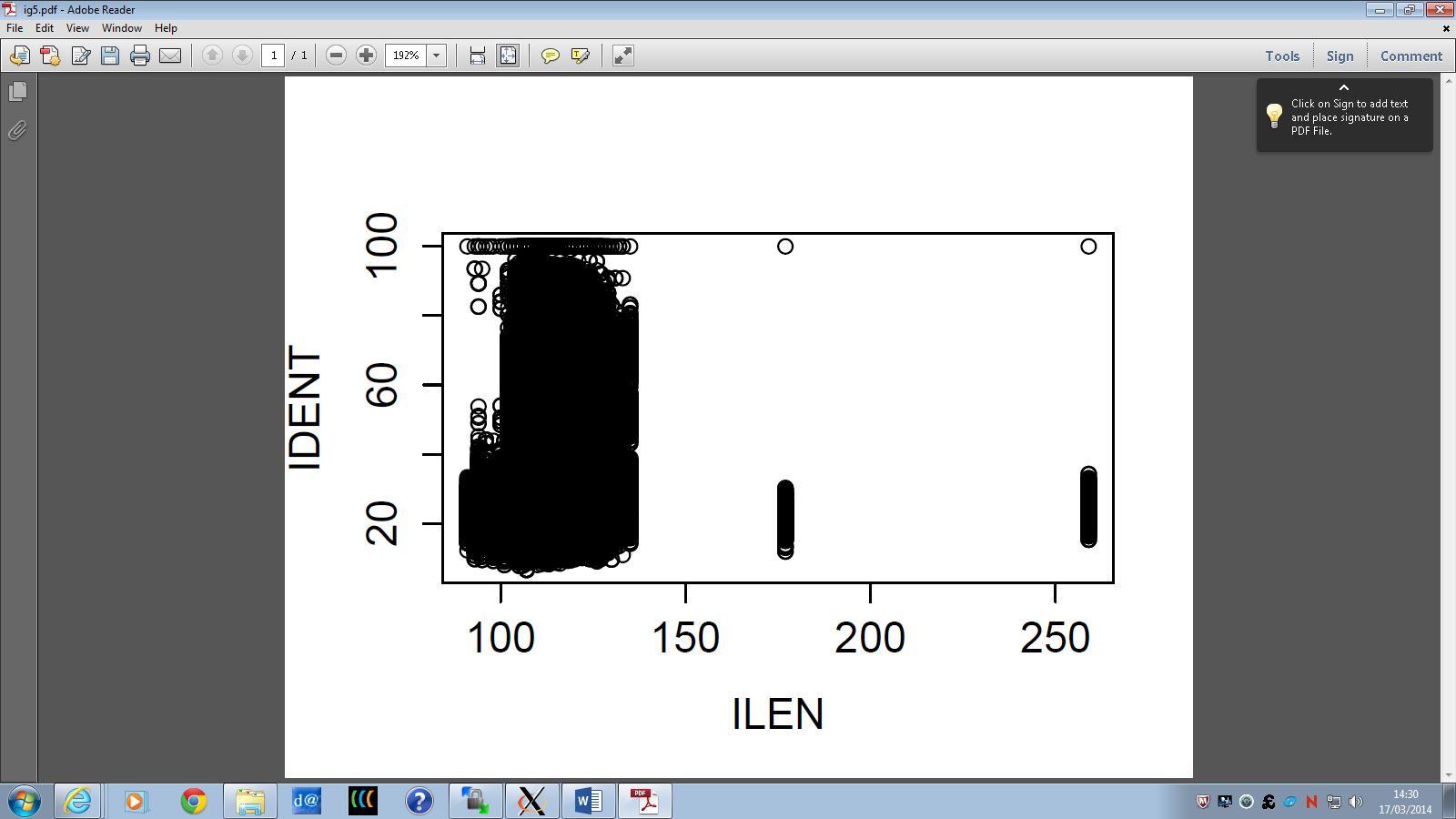
>plot(ILEN,SCORE)#plot6i

Plot4i: ILEN vs MATCH



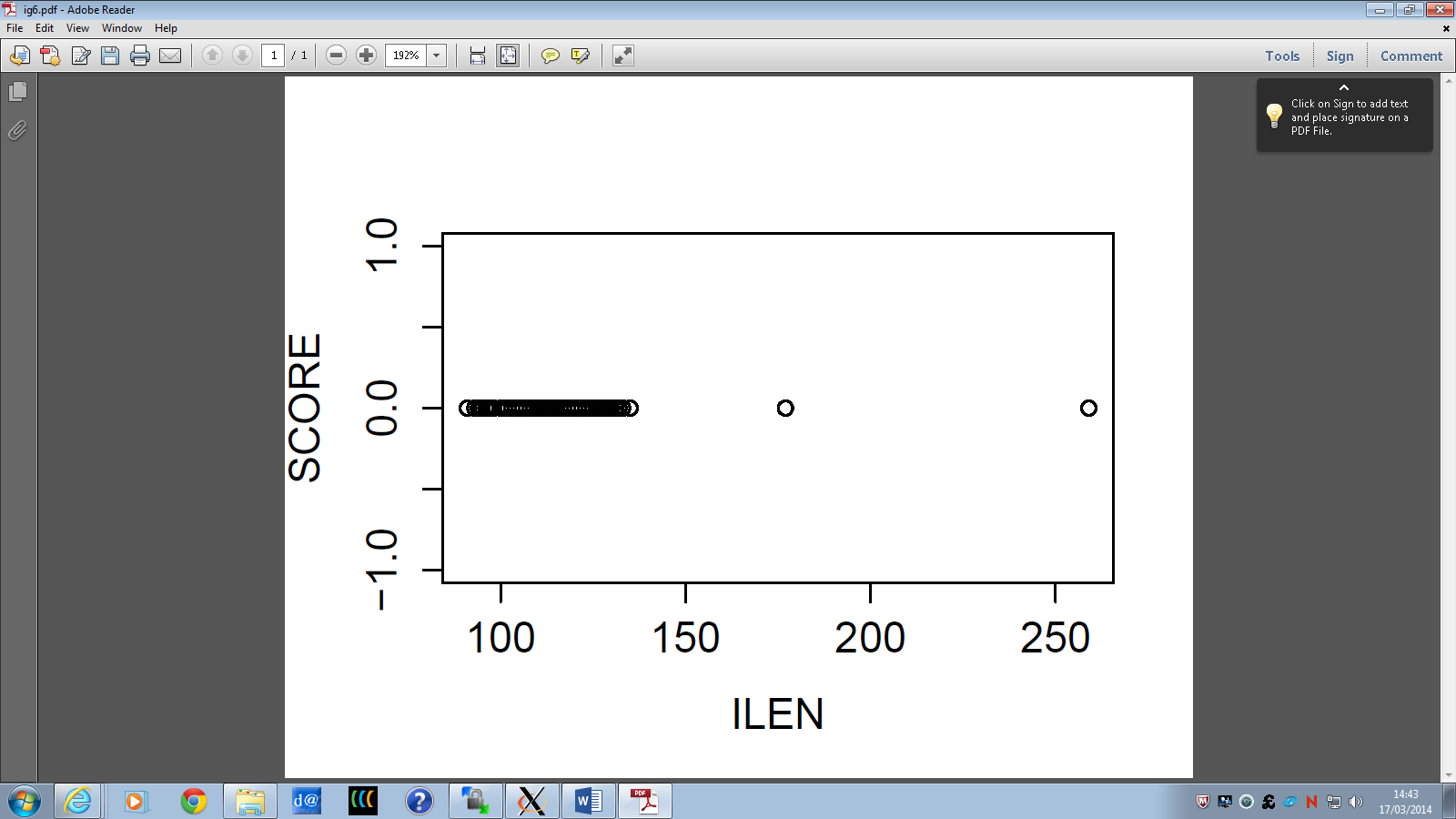
This plot suggests general positive correlation between match score and sequence length. Sequence length is needed to calculate the match score, so this is expected.

Plot5i: ILEN vs IDENT



There does not appear to be any correlation between percentage identity and sequence length. This is what was expected as the two parameters are not dependent on each other.

Plot6i: ILEN vs SCORE



The z-score is always zero in the Ig pairs data, therefore will remain zero for any sequence length.

Serprot Pairs

The following code was used in R to plot the columns against each other:

>x <- serprot\_pairs\_parsed[,"X.IDENT"]

>MATCH <- serprot\_pairs\_parsed[,"MATCH"]

>z <- serprot\_pairs\_parsed[,"SCORE"]

>ILEN <- serprot\_pairs\_parsed[,“ILEN”]

>plot(ILEN,MATCH)#plot4s

>plot(ILEN,X.IDENT)#plot5s

>plot(ILEN,SCORE)#plot6s

Plot 4s=ILEN vs MATCH



Match is a raw match alignment score for the two sequences and should increase as length of the sequence increases. This can be seen in the above plot.

Plot5s= ILEN vs %IDENT(x)



This plot shows that there is not really any correlation between percentage identity and sequence length, therefore they do not rely on each other.

Plot6s = ILEN vs SCORE(z)



The above plot shows very little correlation between z-score and sequence length, suggesting that they do not rely on each other.

Q3. This analysis is just for two protein families. Do you think this will give a good general picture for all protein families?

(3%)

This analysis does give a good general picture for all protein families as they suggest that there should always be positive correlation between the following parameters ;

1. Percentage identity and normalised alignment score
2. Percentage identity and z-score
3. Normalised alignment score and z-score
4. Match score and sequence length

This is because the parameters depend on each other.

The data analysis also suggests that there is no correlation between the following;

1. Percentage identity and sequence length.
2. Sequence length and score.

However there could be exceptions to these correlations. Furthermore , there may be other pairings of parameters that correlate which have not been investigated here.

Q4. Choose some protein sequences of different lengths and different families from Uniprot, then perform pairwise alignment with AMPS and randomisations. Repeat your analysis with this set and discuss.

Two protein sequences were downloaded from the NCBI website:

Sequence 1 = putative FKBP-type peptidyl-prolyl cis-trans isomerase; protein [Oryza sativa Japonica Group] (<http://www.ncbi.nlm.nih.gov/protein/33146997?report=fasta>)

Sequence 2 = protein; K11883 RNA-binding protein NOB1 [Archaeoglobus profundus DSM 5631] (<http://www.ncbi.nlm.nih.gov/protein/284011305?report=fasta>)

Protein sequence 1 is from a species of Japanese rice, whereas sequence 2 is from a species of Archaeoglobus, which is a sulphate-reducing Archaea.

The two sequences were downloaded as fasta files, then both added to Jalview. This groups the sequences together in an unaligned form. This was saved as a PIR file, sequence.pir that could be used for further analysis using AMPS. AMPS is a program that can align multiple protein sequences.

The script sequence.com was written to allow AMPS to be run. This had a gap penalty of 8, a constant value of 8, and a randomisation of (0, 0, 1). The script can be found in the following github repository. <https://github.com/fmacfarlane/Bioinformatics/tree/master/assignment_3>.

AMPS was then ran on the sequence.com file using the command:

$ ./amps<sequence.com

This produces two files, sequence.fasta and sequence.out. The sequence.out file can be opened and the values for specific parameters can be viewed. The fasta file can be displayed in Jalview to show a visualisation of the alignment. I used clustalx colouring to colour the alignment.

This method was then repeated altering the sequence.com file with;

2. Gap penalty= 0, Constant=0 and Randomisation= 0,0,1. (sequence0.com)

3. Gap penalty= 500, Constant=8 and Randomisation=0,0,1. (sequence500.com)

After this had been done each of the above alignments were ran again with randomisation of (100,100,1). This gives values for the z-score of the alignment. The sequence.com files were edited with new values for gap penalty, constant and randomisation.

4. Gap penalty= 8, Constant=8 and Randomisation= 100,100,1.(sequencez.com)

5. Gap penalty= 0, Constant=0 and Randomisation=100,100,1.(sequence0z.com)

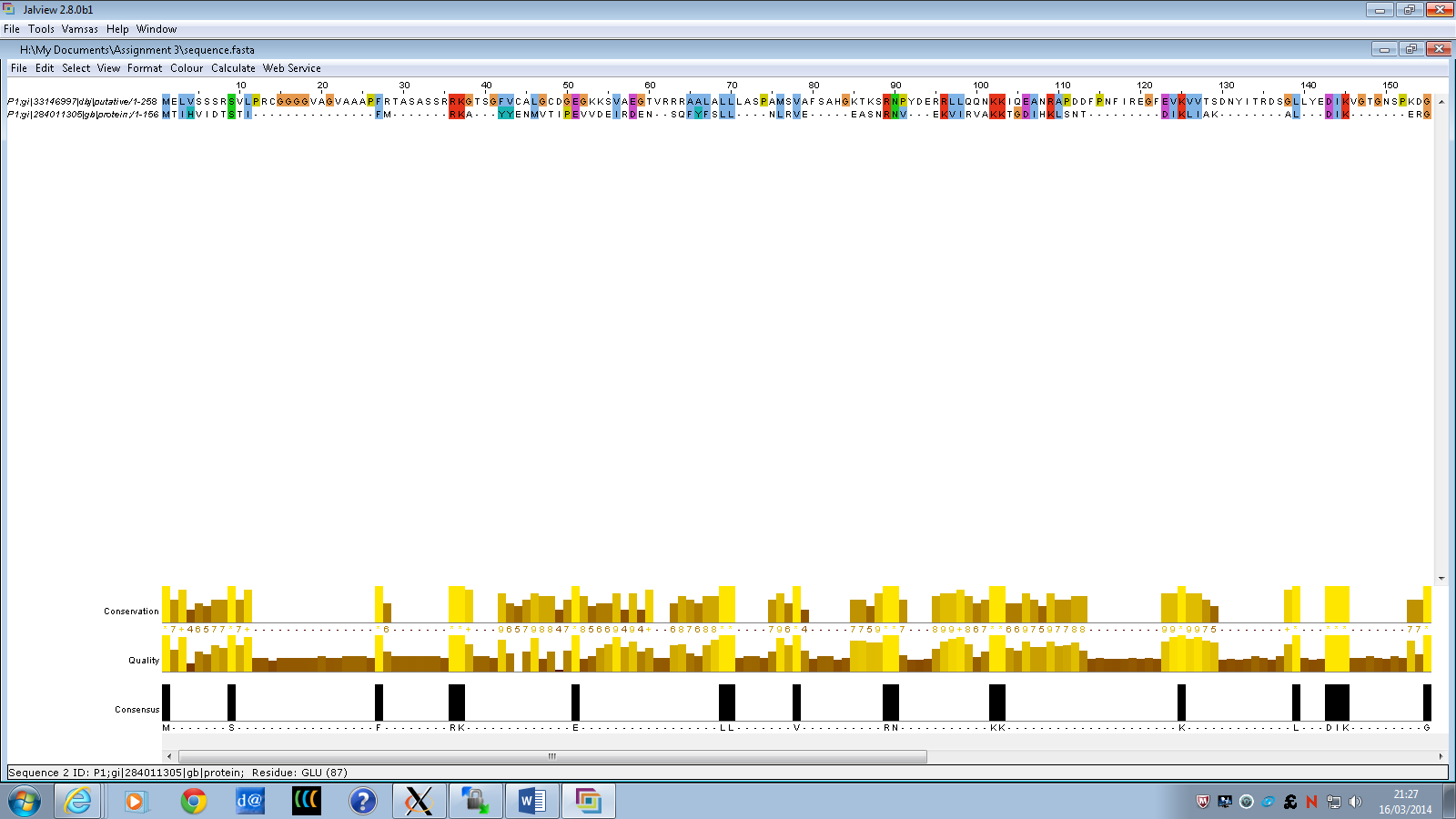
6. Gap penalty= 500, Constant=8 and Randomisation= 100,100,1.(sequence500z.com)

Each of the modified sequence.com scripts can be viewed at the following github repository:

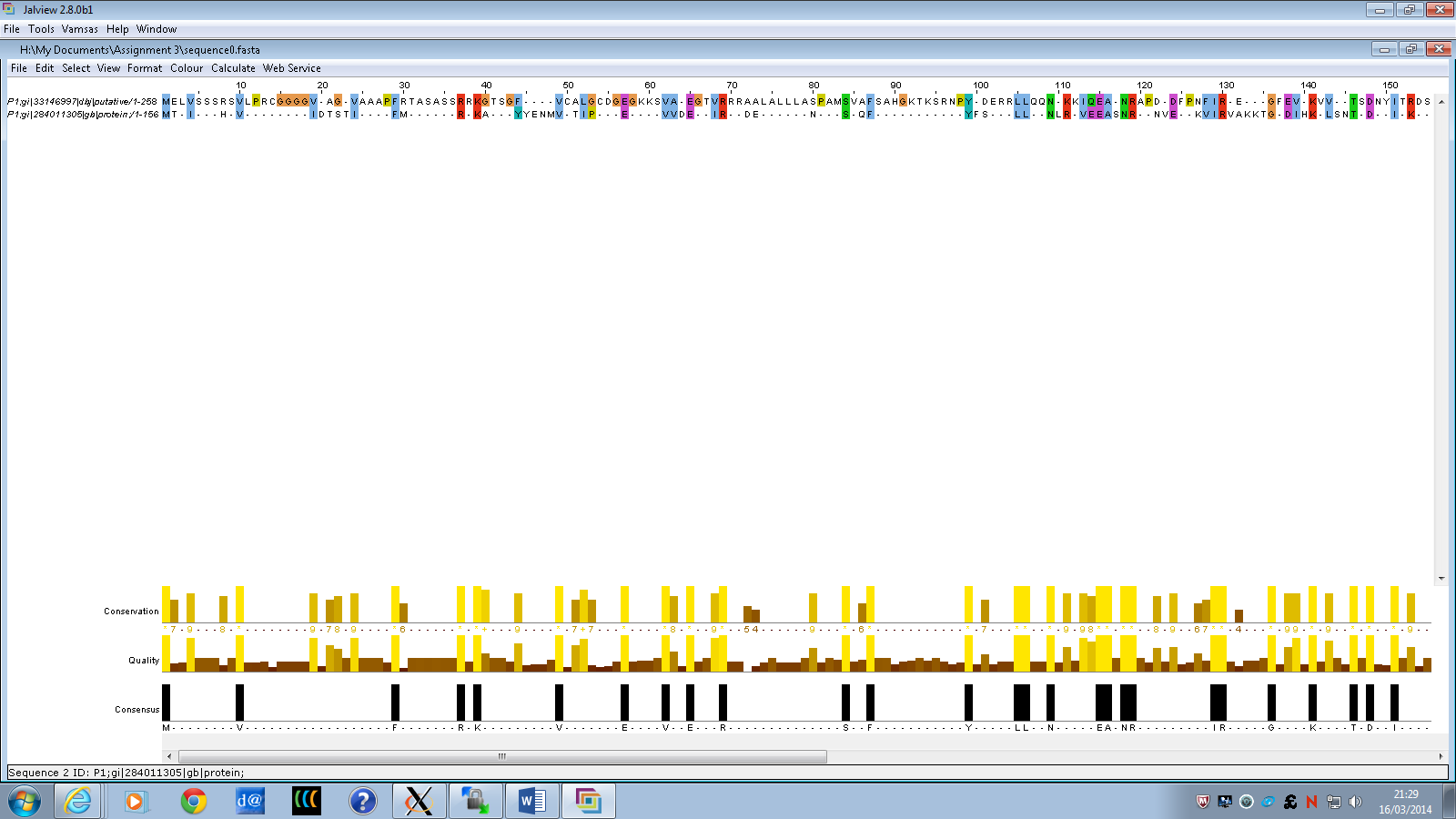
<https://github.com/fmacfarlane/Bioinformatics/tree/master/assignment_3>.

Each of the fasta files were displayed in Jalview with clustalx colouring, screenshots of these are displayed below. Each of the out files were opened and the values for percentage identity (%IDENT), normalised alignment score (NAS) and z-score (SCORE) were noted. The values for these have been inserted into a table below.

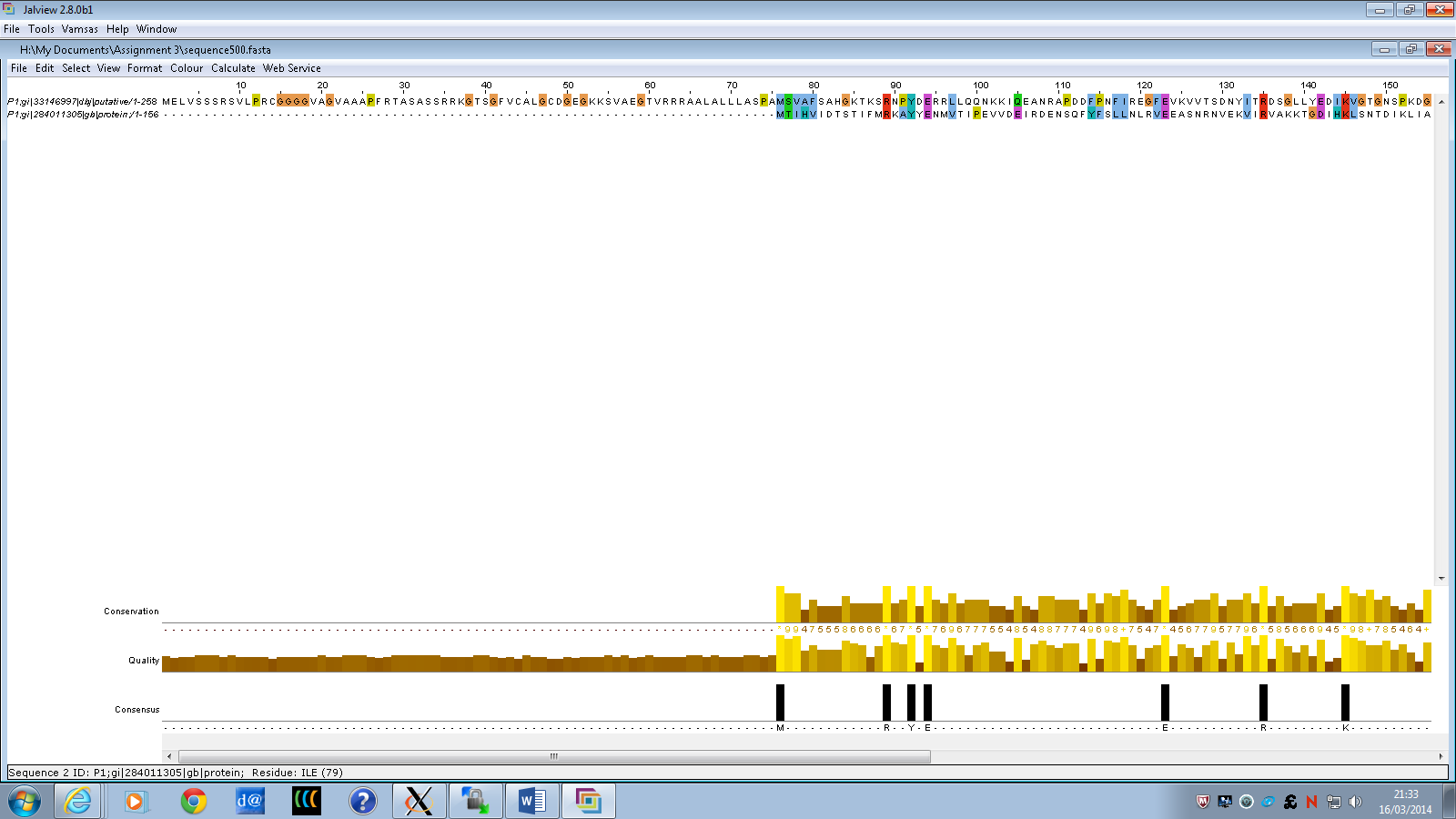
Img1. Alignment of sequences after amps has been run., random=0,0,1 gap=8 c=8:



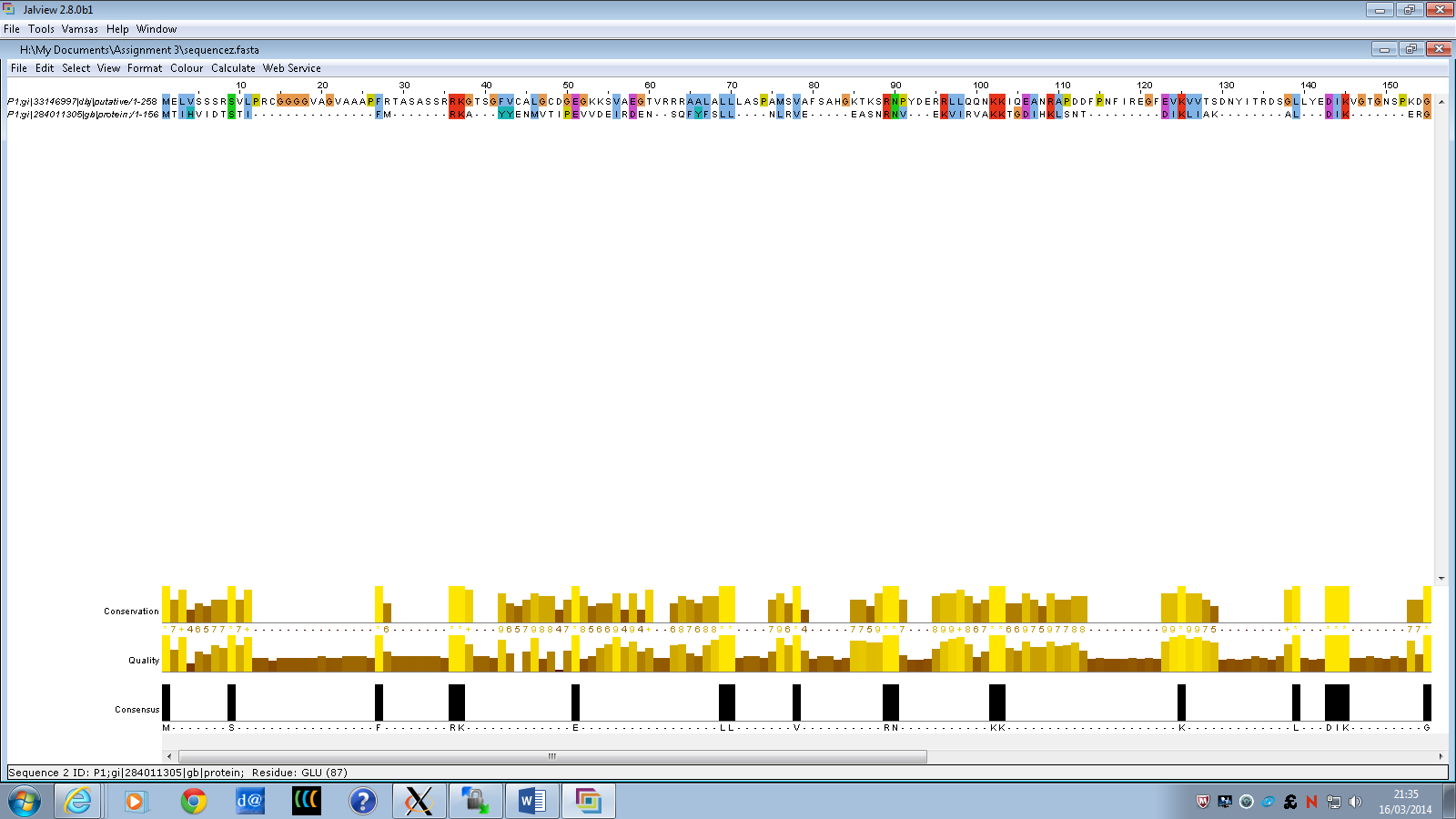
Img2. Randomised gap penalty = 0 constant = 0 , random=0,0,1:



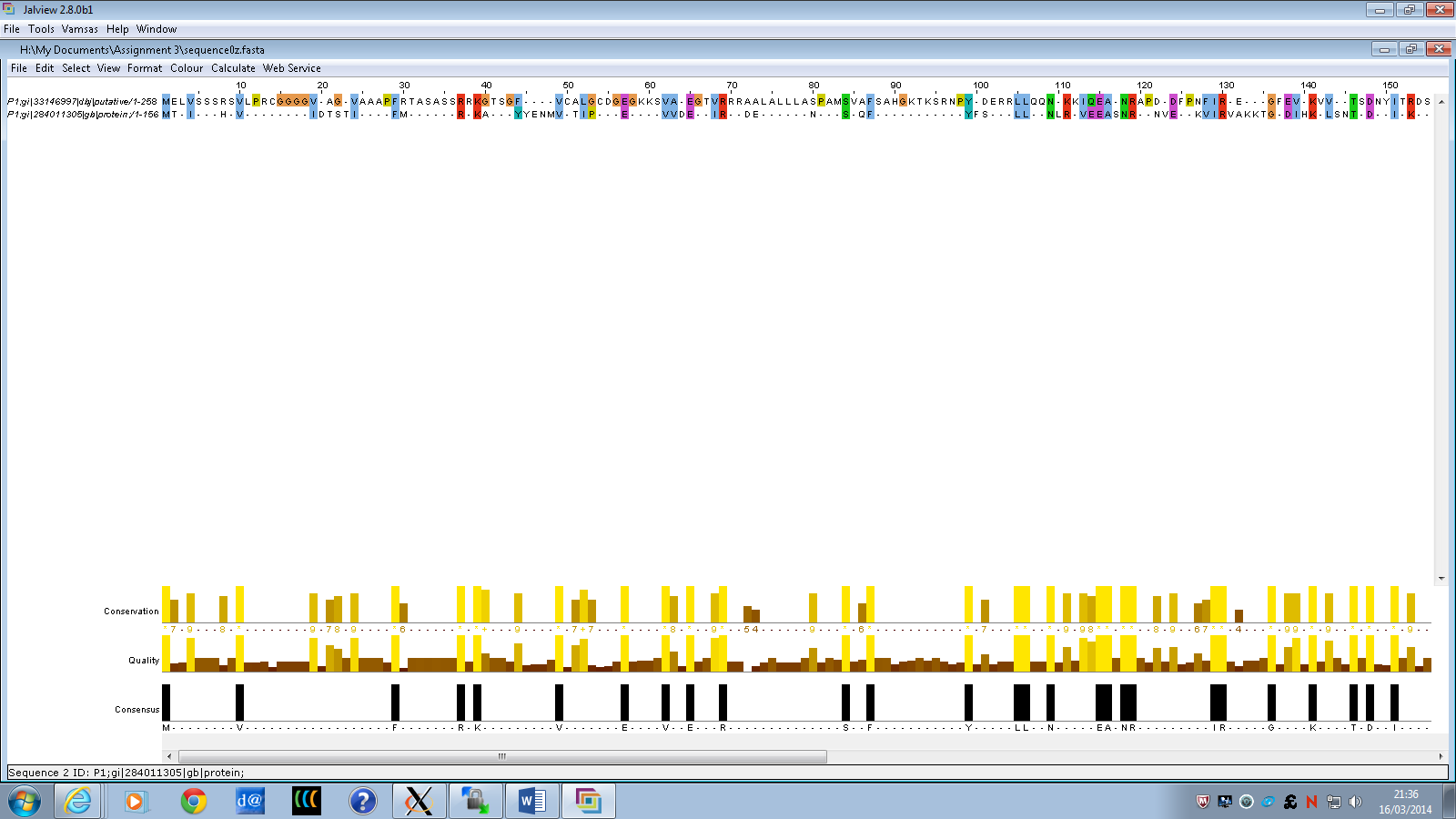
Img3. Randomized gap penalty=500 constant = 8 random=0,0,1:



Img4. Randomised z- score, gap penalty=8, constant=8, random=100,100,1:



Img5. Randomized z-score gap=0, c=0 random=100,100,1:



Img6. Randomise z-score gap=500, constant=8, random =100,100,1 :

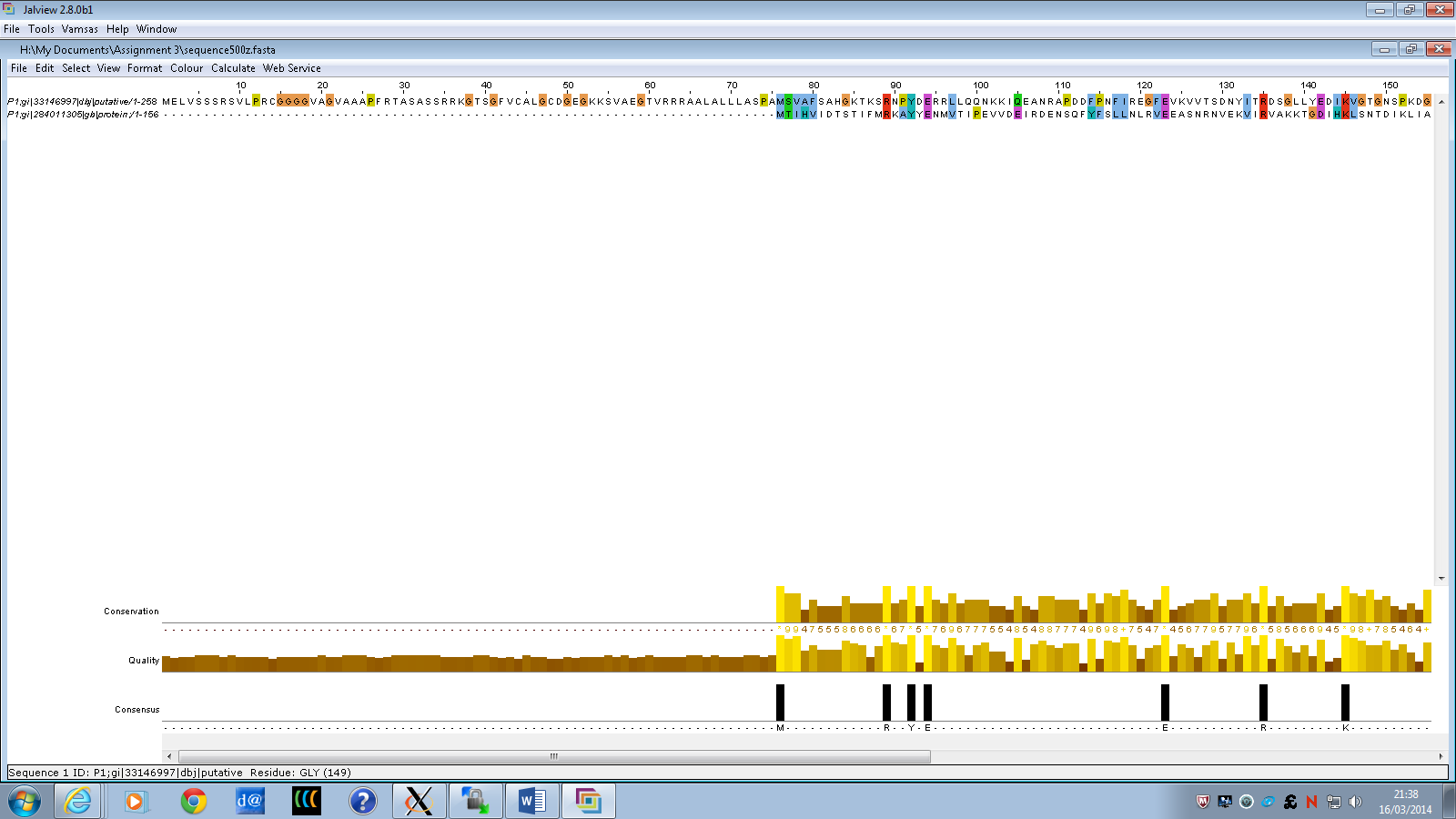


Table of statistical values:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 |
| %IDENT | 22.58 | 51.3 | 8.97 | 22.58 | 51.3 | 8.97 |
| NAS | 844.52 | 352.17 | 784.62 | 844.52 | 352.17 | 784.62 |
| Z-Score | 0 | 0 | 0 | -1.82 | -1.15 | 1.71 |

A z-score of greater than 6 suggests that the alignment is correct over most of its length. None of the alignments above have a z-score of over 6, suggesting that the alignment is not reliable. Alignment 6 has the highest z-score value, so is the most reliable alignment to choose in this case. The percentage identity value for this alignment is low, 8.97%. For a percentage identity to be relevant it has to be between 0-20%. This suggests that the two sequences are not related and the percentage identity score is just coincidental, rather than a phylogenetic relationship. It would be surprising if the two species were related as they are from different families and are of different lengths. The Jalview display of Alignment 6 also shows that there is a large gap at the start but not many gaps after that. This supports the suggestion that alignment 6 is the strongest alignment of the protein sequences.