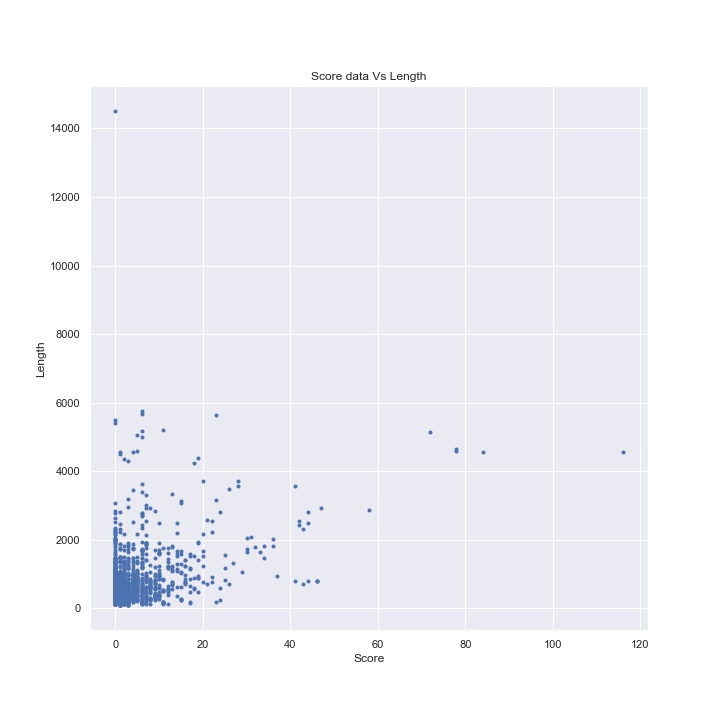
BIOINFORMATICS

Score All(Size=2768)



Data plotted between SCORE of disulfide bond and LENGTH of the sequence. Most of the data is compressed in the small section of SCORE between 0 and 20 and LENGTH 0 and 2000. So to explore more we need to dig more inside.

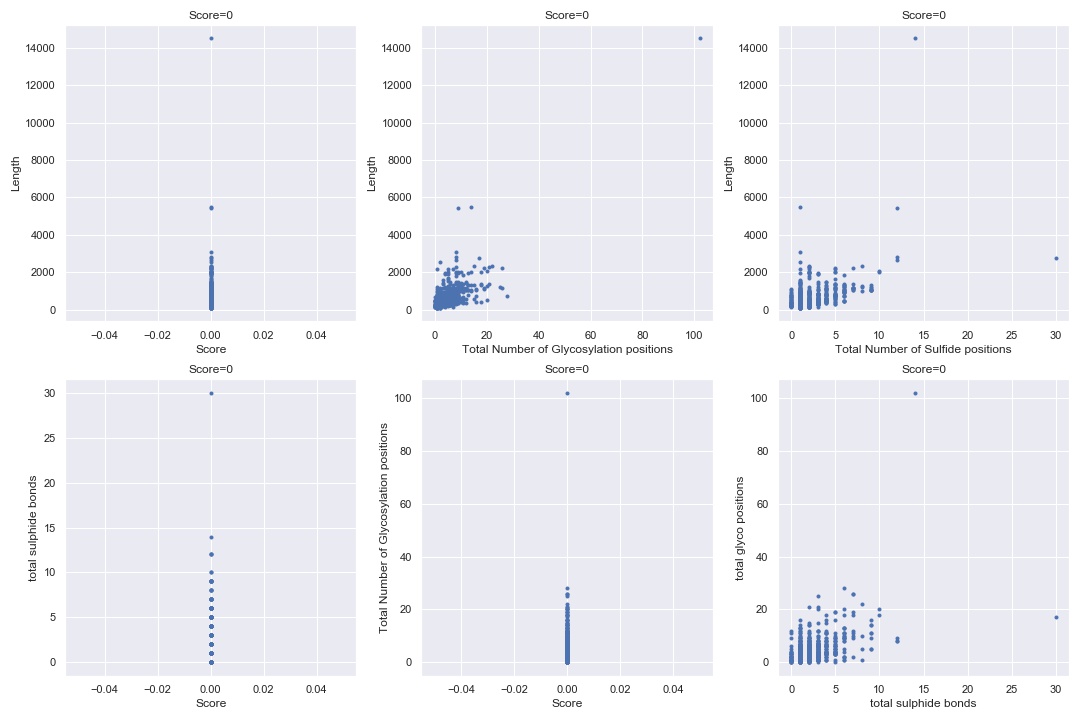
Clustering Manually [Using Python]

Four Clusters

* Score=0
* Score>0 and Score<=10
* Score>10 and Score<=20
* Score>20

Disulfide bond Frequency and Glycosylation Frequency also Visualized

Score=0 (Size=1526 ~ 55%)



Graph1 shows Maximum Length of Sequence is around 3000 with two outliers near 14000 and 6000.

Graph2 shows there are 2 proteins with Length near 6000 and have Glycosylation positions less than 20 and 1 with Length near 14000 have Glycosylation position more than 100.

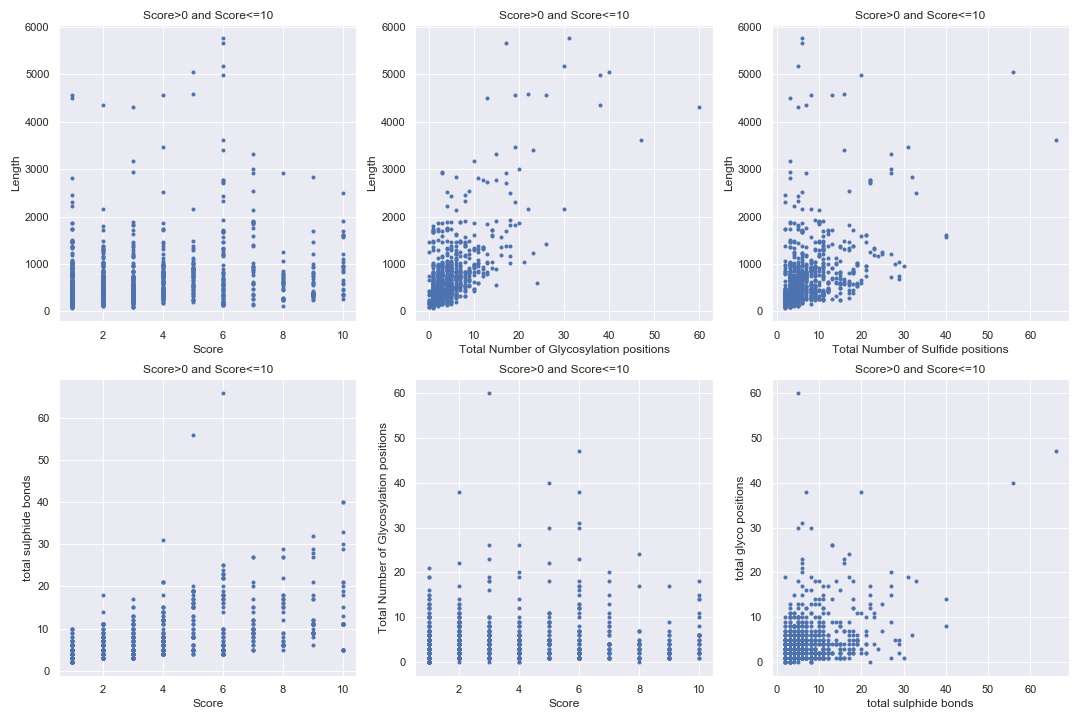
Graph3 shows 1 more outlier with Sulfide Positions 30.

Graph4 similarly shows property like graph 3.

Graph5 similarly shows property like graph 2.

Graph6 shows scatterplot with 2 outliers and rest the data coincides at one place.

Score>0 and Score<=10 (Size=1086 ~ 39%)



Understanding these graphs is very complex as data distribution is wide.

The orange indicates an outlier with score=6, Sulfide bonds >60 and glyco position>45. While one more outlier with glyco positons=60 and sulfide pairs<10 and score=3.

Score>10 and Score<=20 (Size=105 ~ 3.7%)

A close up of a device

Description automatically generated

Score>20 (Size=51 ~ 1.8%)

A picture containing white

Description automatically generated