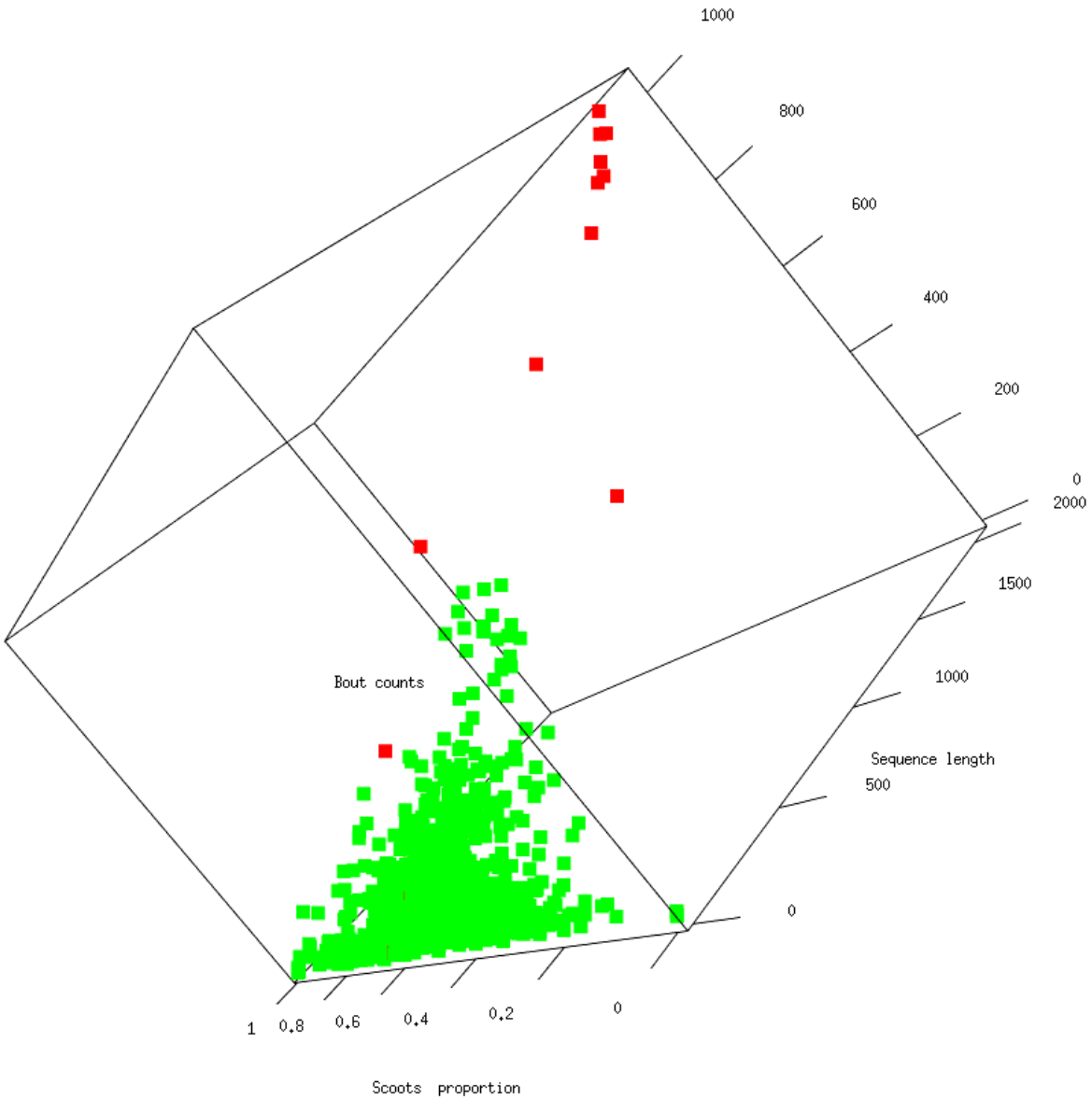
UPDATE 27.01.2017

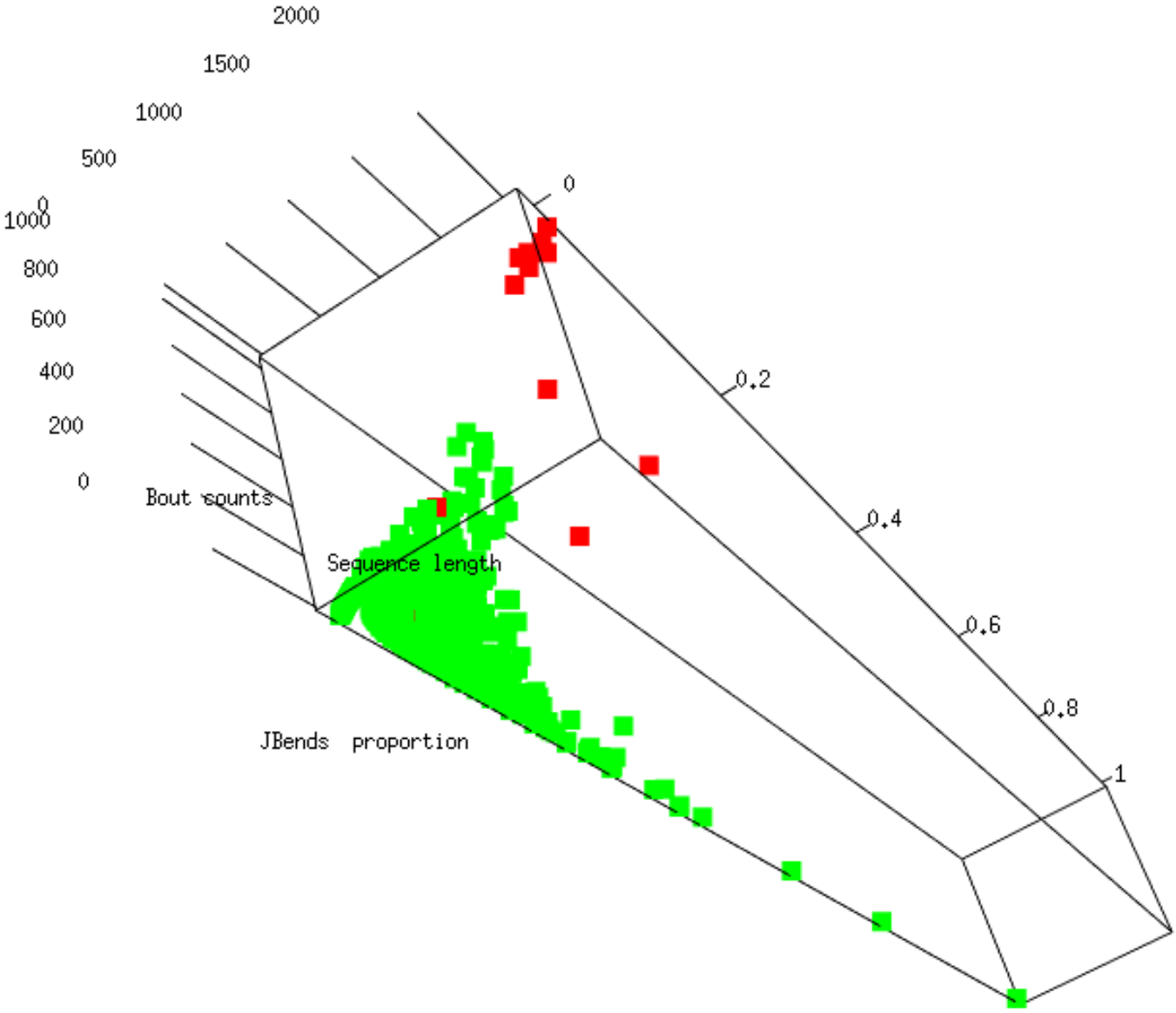
The past week I have been trying to find the best way to identify the outliers, so I can extract them. Fredrik has given me information about the individuals and the corresponding time-frames, where he was sure that the tracking software produced errors. These are 11 cases within the Dark Apo Low condition for Aripiprazole and Cariprazine. Based on the attributes of those action sequences , I tried to identify error action sequence in the entire dataset. One thing to note is that these cases were mixed, so control cases and different drug dosages. But nevertheless, it appeared to group/separate into normal and outliers well.

I had looked at the action sequence length, bout count and the over all turn proportion. The error action sequences were much longer, had a greater bout count, so lots of very short bouts and the turn proportions were different from the normal action sequence.
Here are 3D plots for all turn proportions, plotting sequence length, bout count and turn proportion, if you run R code, you can actually hold and rotate the graph yourself, but since you will probably not run R, I did it for you and saved it as gif(folder “plot_rotations”). Red are outliers, green are normal.

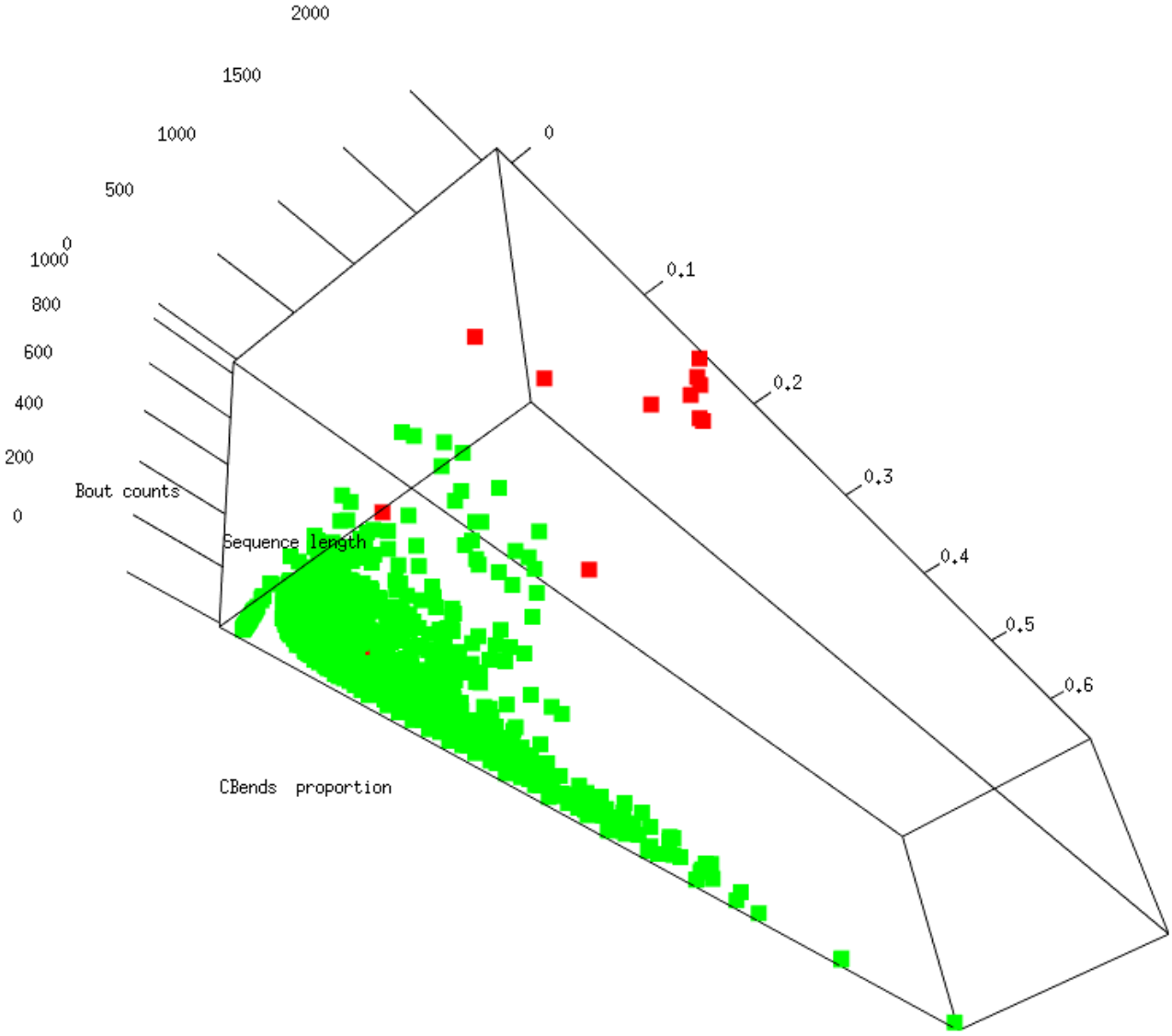
Scoots



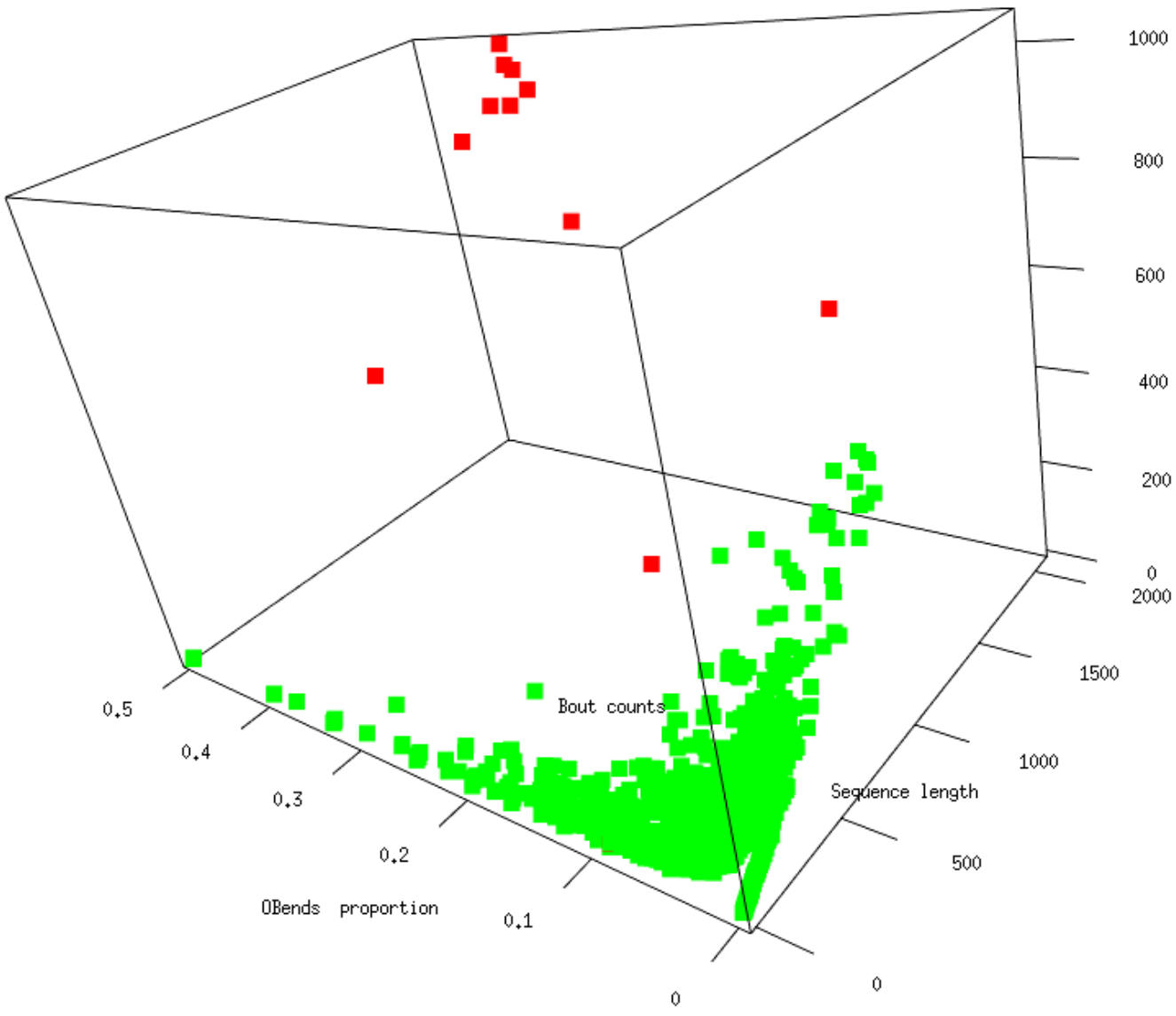
JBends


2000
1500
1000
500
0
1000
800
600
400
200
0
Bout counts
Sequence length
0
0.2
0.4
0.6
0.8
1
JBends proportion

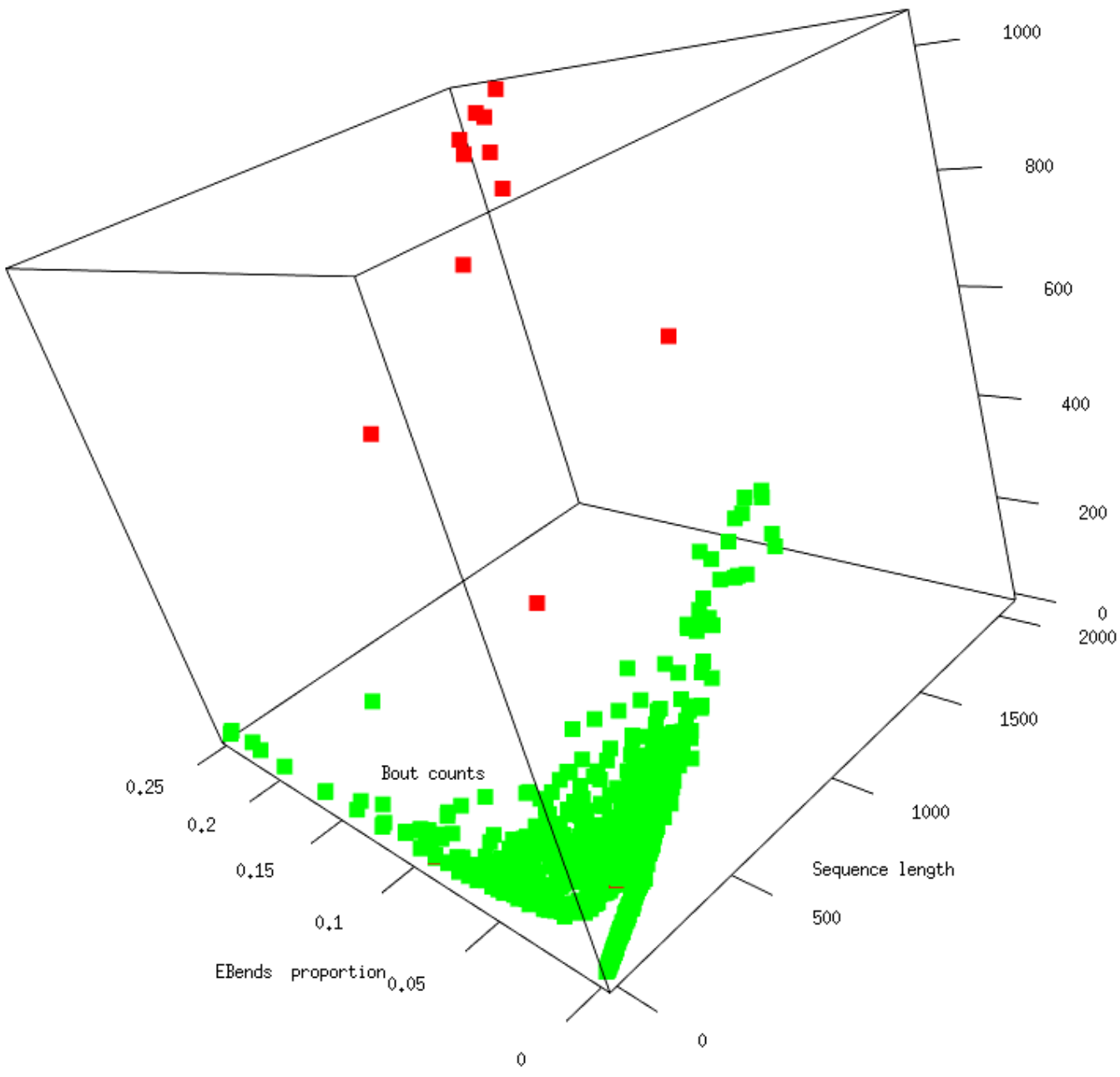
CBends


2000
1500
1000
500
0
1000
800
600
400
200
0
Bout counts
Sequence length
0
0.1
0.2
0.3
0.4
0.5
0.6
CBends proportion

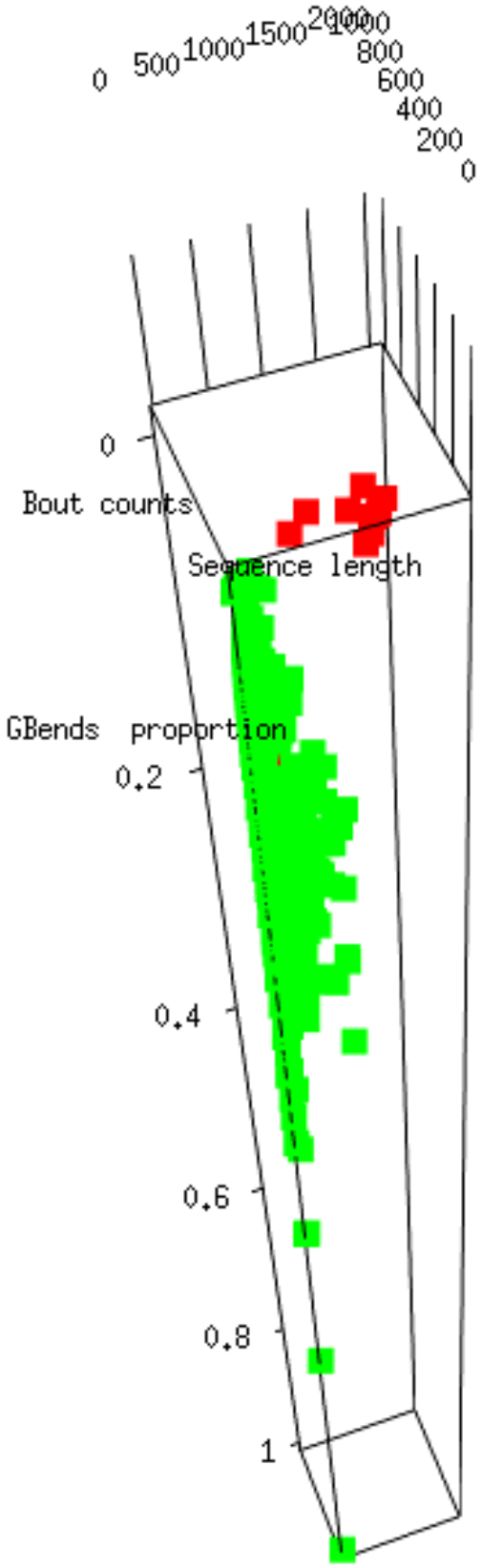
OBends


1000
800
600
400
200
0
2000
1500
1000
500
0
Sequence length
0.5
0.4
0.3
0.2
0.1
0
OBends proportion
Bout counts

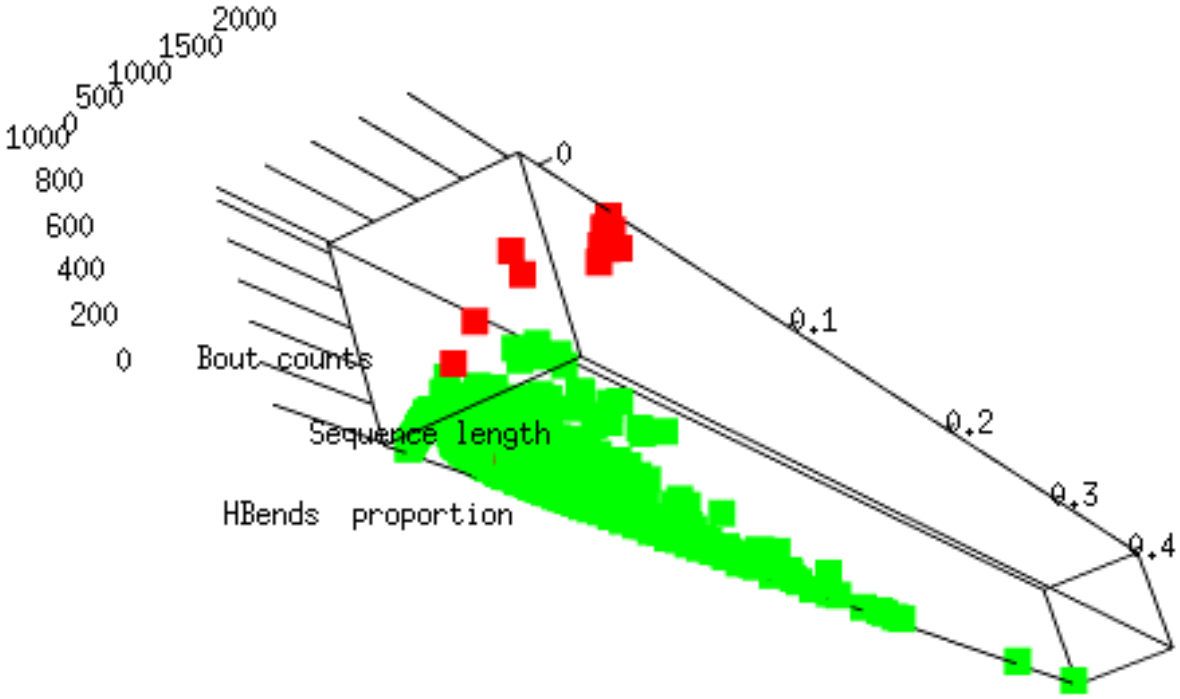
EBends


1000
800
600
400
200
0
2000
1500
1000
500
0
Sequence length
0.25
0.2
0.15
0.1
0.05
0
EBends proportion
Bout counts

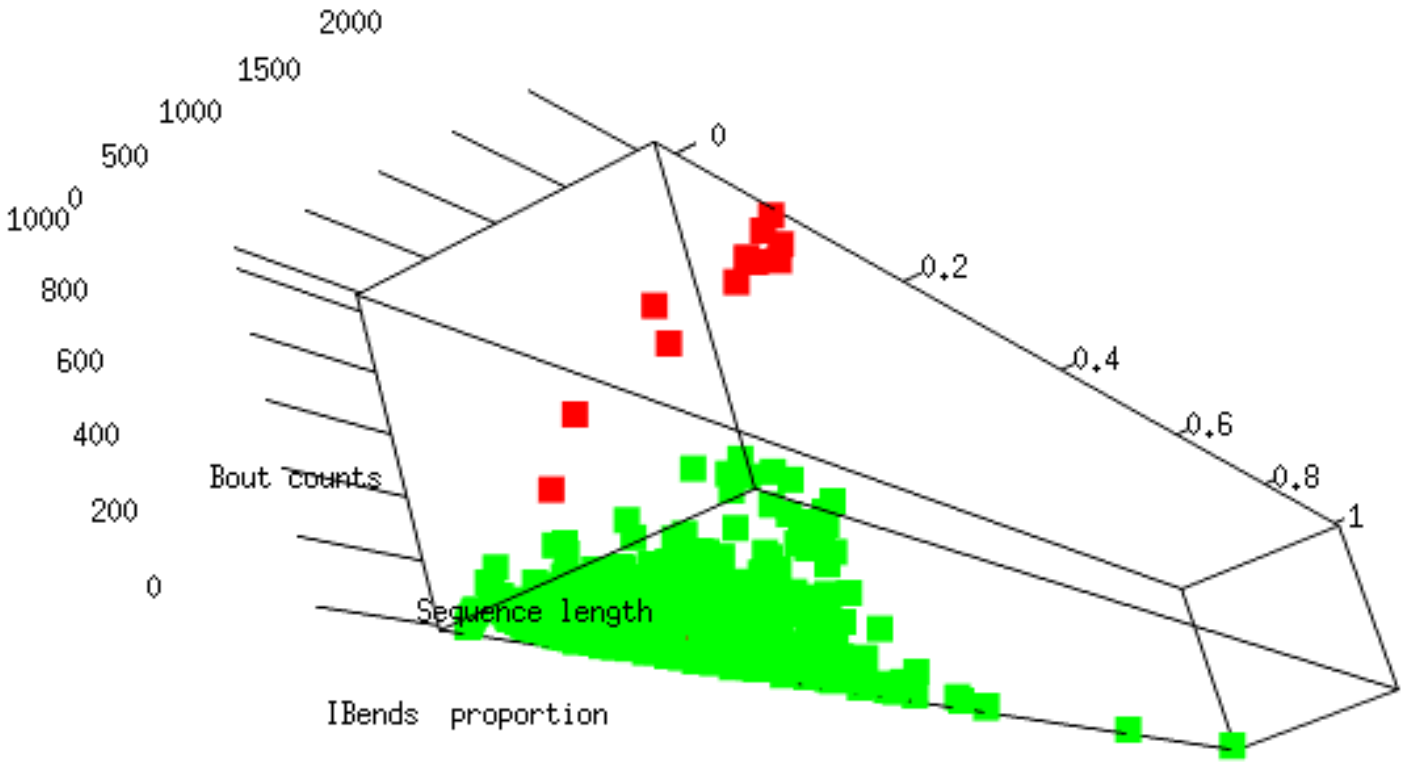
Gbends


0
500
1000
1500
2000
1000
800
600
400
200
0
0
Bout counts
Sequence length
GBends proportion
0.2
0.4
0.6
0.8
1

HBends


2000
1500
1000
500
0
1000
800
600
400
200
0
Bout counts
Sequence length
HBends proportion
0
0.1
0.2
0.3
0.4

IBends


2000
1500
1000
500
0
1000
800
600
400
200
0
Bout counts
0
0.2
0.4
0.6
0.8
1
Sequence length
IBends proportion

This is only within DarkApoLow, within Aripiprazole and Cariprazine. I am not sure if those are the only outliers within Aripiprazole and Cariprazine(looks like it). But in the entire Dark Apo Low there seems to be a bit more of similar points, I plotted including the rest in DarkApoLow and then I also plotted for all the rest of the cases in the other conditions, where its apparent that the cases are too different between the conditions, so most probably outlier correction will have to be done inside each condition.

So I modeled the data of normal and outliers with a random forest, using only the data from DarkApoLow, within Aripiprazole and Cariprazine. I had to substantially undersample the oversized normal group(random undersampling) and oversample the undersized outliers grouped(SMOTE). Then I used the model to classify the rest of the action sequences into normal or outlier in the rest of the DarkApoLow and just for comparison, in the rest of the conditions also.

So for normal cases I used only the rest of the DarkApoLow Aripiprazole and Cariprazine, and I excluded the zero length bouts. So the number of normal cases was 1222 and the number of outliers 11.
I undersampled the normal group to around 800 and oversampled the outliers to around 400.
I then divided the set to learning and testing and used random forest with 2000 trees.
Details:

Learning set:
```
proportion of 0: 0.3912214
proportion of 1: 0.6087786
```

Testing set:
```
proportion of 0: 0.3374656
proportion of 1: 0.6625344
```

Results:
```
#> table(predictData,unlist(AllTesting$Class))

#                       AllTesting$Class
#predictData                  0    1
#0                            245  3
#1                            0   478


#class. acc.: 0.9958678
#sensitivity: 0.993763
#specificity: 100
#precision: 100
```

I had plotted the predicted unknown classifications, along with known. Green is normal, red is known outlier, blue is predicted outlier. Folder “all_plots” has all the plots, so you can compare, “limited” is when the data is limited only to DarkApoLow Aripiprazole and Cariprazine , “allDarkApoLow” is the entire DarkApoLow condition and “allConditions” is entire set of 6 conditions.
As I mentioned, it is apparent in the plots, that with cases from only DarkApoLow the rest of the conditions are probably not predicted well.
I am sending the list of all predicted outliers in the DarkApoLow and in the rest of the conditions.

I also checked the presence of any odd motifs, like sequential rare turns, but with the amount of sequences I got form Fredrik, nothing comes out as significant, still sequential scoots are the most common. From what I had observed, the bouts in the outliers are so short, usually just one turn, so motif search is a bit impaired anyway.