UPDATE 17.03.2017

After removing the outliers which were produced by the tracking software I had rerun some of the previous analyses. I had noticed that in some cases the data will still look like there are some outliers, not such that were produced by the tracking software, but maybe some exceptions that the fish did, like for example just one bout with unusual turn proportions and I am wondering if we should exclude those as well. Maybe look at some variables like the turn proportion per bout length, creating a normal distribution and excluding the lower and upper 5%?

The only sensible analysis from the previous analyses, besides the cluster analysis and the statistical analysis of turn proportions that I am already working on, I think that only the motif analysis would be good to continue working on now, others were just exploring data or were related to the other two analyses.

I had a better look at the plots now that the data was cleaned and extracted the motifs for all the data, so the control and 3 different dosage for 12 drugs, all in 3 healthy conditions and 3 disease conditions.
I am not yet sure how we are going to define and estimate the significance of the motifs, but it seems as if motifs up to and including length 5 are sensible to look at. Beyond that length, the counts become very small and could be just random occurrences. The data and all the plots are on git in the folder “simple_motif_search”.

So I will have to study this and come up with a way to assign significance to the motif occurrences. We already know that the probabilities for the turn proportions are not equal, so I am not sure how appropriate would it be to multiply 1/8 for the word length to get the probability of a certain motif and then do a binomial test with n being the total number of words of that length and k being occurrences of a specific motif.
And the turns dont seem to be independent from each other. Looking at the plots, after the most common motif of all scoots for any word length, the next few most common motifs will always start with a different turn type from scoot, most commonly G bend and that turn is then followed by scoots. Since the transition from scoot to a different turn type is less common, might be that fish start with a different turn type and then continue with scoots.

So for now, besides doing the cluster analysis and the statistical test for the turn proportions, I will look at the transitions for the turn types, so focusing on 2 letter motifs, while also looking at the first turn type of the bout. And I will study ways to estimate the statistical significance for the motif occurrences. There are many papers and tools for estimating the statistical significance for motif occurrences, but usually they work with oligos in a genome, focusing on associations with biological functions, like for example finding motifs that represent binding sites. So I will have to understand those really well in order to adapt them for the action sequences.