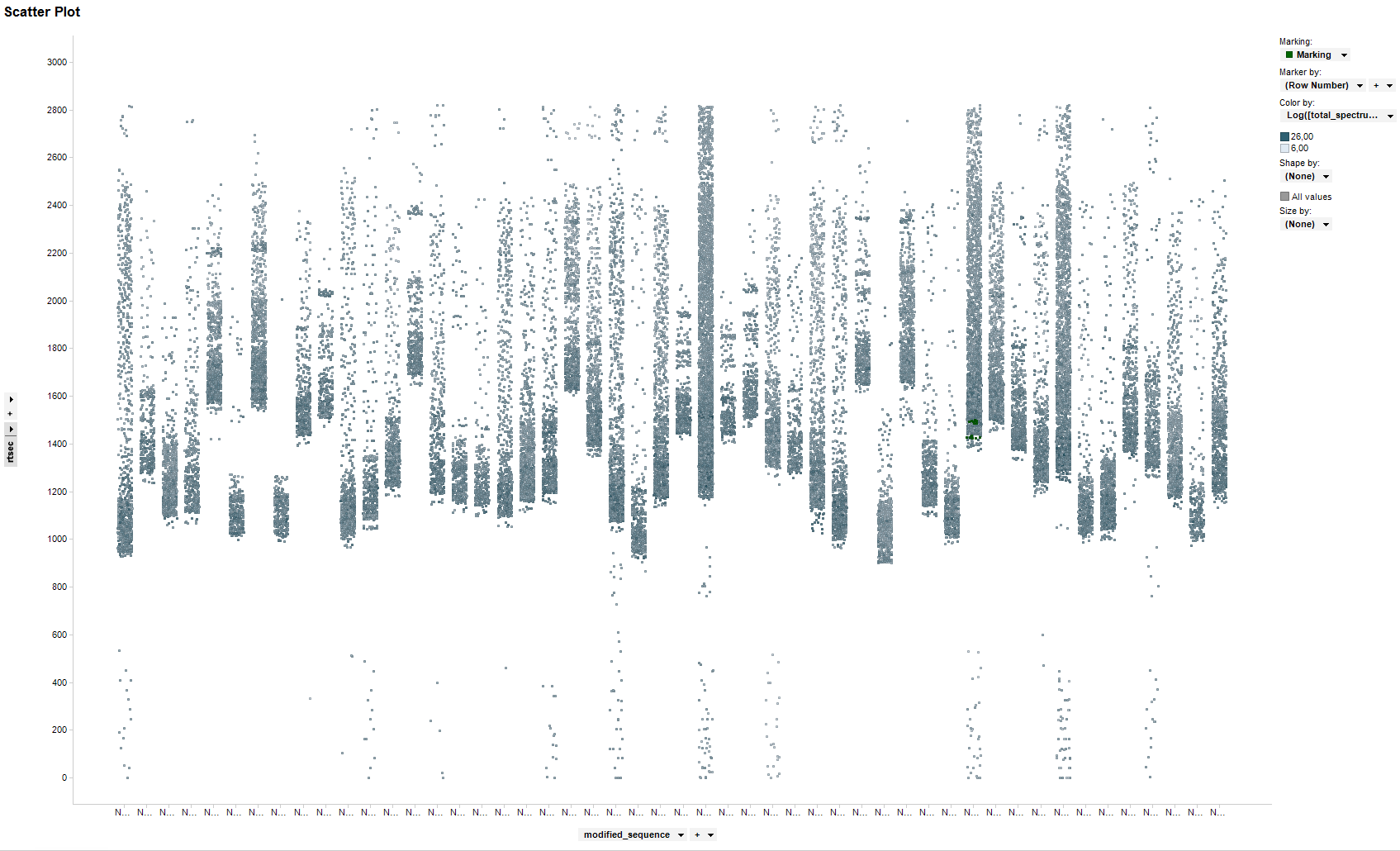
Problem: for retention times, we often have a good idea about the starting point, but it is unclear when a peptide stops to elute.

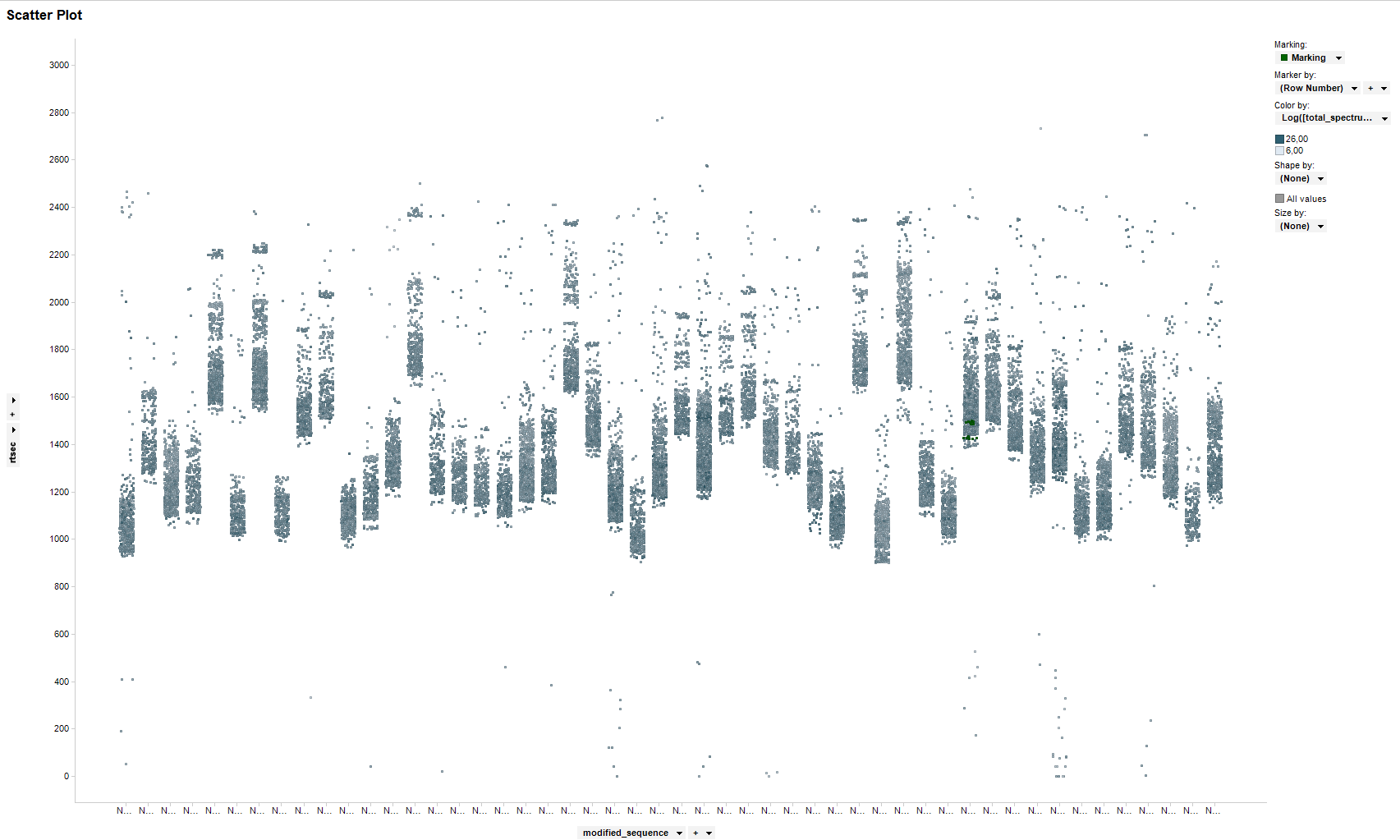
In the context of the Inspector project, the purpose of retention time prediction is to predict more efficient SRM (selected reaction monitoring) design : if we know a peptide elutes at some point, we can tell the scoring algorithm to consider this peptide within this tiny retention time frame. Currently, this process is already somehow done, “manually”, but we want to make this process automatic.

One of the “choke points” in this automation process is the very big tail. Is it related to some projects, to some peptides/proteins, or even both ?

Keeping only the most intense spectrum (MS2 related) of a LC-run was a bit harsh, instead we created an index : 1 for the highest intensity of a peptide in a run, 2 for the second, etc…

It appeared that most of the time, peptides are not identified more than 3 or 4 times per LC-run, and in some cases are identified dozen of times. By filtering out identifications where the peptide has already been identified at higher intensities, it reduces greatly this “big tail” problem. Allowing 4 identifications of one peptide per LC-run seems a good compromise.





Those 2 graphics show the retention times of the top 50 peptides (number of projects related). A lot of them have a very long tail, and the view gets better when filtering over-repeated identifications.

Comments after discussion with An on August 20th 2013 :

* Pump of Rita and Linda is faster. Bigger gradient (less precise) than Vanessa. Emanuella has the same pump as Vanessa, should be more precise but has been widely used only since May 2013.
* Emanuella doesn’t have any ion-trap, everything is measured in the Orbitrap and HCD : higher accuracy and resolution.
* N-terminal COFRADIC tails less than shotgun. Unfortunately, the point of rt prediction is for shotgun experiments, where complexity of the sample are very high. We suspect that higher complexity of sample induces more tailing. Have a look in the description column of the table “identification”: it will tell what kind of protein is analysed: human protein should be more complex (thus with more tails) than yeast.
* There was indeed a big change one year ago with Vanessa : use of 2 columns instead of 1. After a short period of trial, experiments became way more stable. However, the retention time seems to be slightly increasing over time. An says it might be because the pump becomes slower. To be sure about that, look up the ENOLASE DB on Vladimir : the same protein is regularly analyzed there, we should see the same kind of retention time increase. \*Wait, wouldn’t that be a way of normalizing retention times?? \*

What seems to be usually done with retention time predictions : we make RT depending on many factors, and then we have a prediction, which can be more or less precise.

What if we try to predict a probability that the retention time is between X-n and X+n seconds? The thing is that probably what will happen is that the model will retain the most important features. I’m doubtful it will work anyway.

22/08/2013 : with support vector regression, this is done somehow, “soft margin” means that if the difference between predicted retention time and observed retention time is less than a threshold sigma, then the penalty in the optimization algorithm is 0. The algorithm takes into account that there is noise in the data.

23/08/2013 : update following discussion with Lennart.

First, description of N-terminal COFRADIC:

* Acetylation of proteins
* Digestion with Trypsin
* First LC-Run without any MS : output is fractionated into buckets containing rt frames of roughly 5 minutes
* TNBS added to each fraction : it will react with N-terminal and induce a big hydrophobic shift
* Another LC-Run with MS, peptides eluting in the original rt frame will be the acetylated N-terminal of the protein

Given that the second LC-Run only contains N-terminal of proteins, the complexity of the sample is greatly reduced.

Always keep in mind that the rt prediction for Inspector has the purpose of designing better SRM.

Trying to bring closer together two sets of run which are obviously different, I noticed that the shift had a tendency to grow over time. If that’s true, it means the error in rt prediction should grow over time. It may be very specific to the sets of run >> focus on higher quality data, and see if there is still a relationship between rt and error.

About big tails : they are specific to some kind of projects >> does it bother the retention time of other peptides ? Hard to tell.

Next step is trying to apply some rt prediction algorithms and see how the data is handled.

And try to answer those questions: does the complexity influence the error? Difference between N-terminal COFRADIC and Shotgun? Difference between E. coli, Yeast and Human?