## May 15-17, 2012

Trimmed sequences to putative initiating methionine then ran SpiderP to call signal peptide, propeptide, mature boundaries.

<http://www.arachnoserver.org/spiderP.html>

Organized results according to signal peptide, propeptide similarity as well as Cys structure patterns.

Subsetted 8 Cys groups according to patterns of

CCC

CECCG

CC,CYIC

CC,CXC,CXC

Looked into lower scoring groups 22-29.

Observed that Group 23 looks like a fusion toxin, misassembled?

* + 12 Cys, 2 pairs of adjacent CC’s, maybe 11 Cys in mature form suggesting a likely active dimer
  + perhaps 2 concatenated 6 Cys toxins
  + Present in 2 libraries so more likely real than misassembly

Assembled a FASTA file of curated sequences

## May 10, 2012

Aligned protein groups (**1,5,8,12,13,15,20,21**) not containing a CCC motif and 8 Cys.

37\_1680.3d.comp348\_c0\_seq1\_F5\_R3|spider.q17.comp1237\_c0\_seq1\_F2\_R3

37\_1680.3d.comp516\_c0\_seq1\_F2\_R2|spider.q17.comp1532\_c0\_seq1\_F3\_R3|AS390

37\_1680.3d.comp410\_c0\_seq1\_F2\_R3|spider.q17.comp1386\_c0\_seq1\_F2\_R3

spider.q17.comp2075\_c0\_seq1\_F2\_R2|37\_1680.3d.comp820\_c0\_seq1\_F2\_R1

37\_1680.3d.comp454\_c0\_seq1\_F4\_R5|spider.q17.comp1469\_c0\_seq1\_F4\_R5

37\_1680.3d.comp454\_c0\_seq2\_F4\_R5

spider.q17.comp1557\_c0\_seq1\_F3\_R1

spider.q17.comp1562\_c0\_seq3\_F5\_R4|spider.q17.comp1562\_c0\_seq1\_F5\_R1

1,8,12 very different from the rest by sequence

Aligned protein groups (**5,13,15,20,21**) not containing a CCC motif and 8 Cys.

37\_1680.3d.comp516\_c0\_seq1\_F2\_R2|spider.q17.comp1532\_c0\_seq1\_F3\_R3|AS390

37\_1680.3d.comp454\_c0\_seq1\_F4\_R5|spider.q17.comp1469\_c0\_seq1\_F4\_R5

37\_1680.3d.comp454\_c0\_seq2\_F4\_R5

spider.q17.comp1557\_c0\_seq1\_F3\_R1

spider.q17.comp1562\_c0\_seq3\_F5\_R4|spider.q17.comp1562\_c0\_seq1\_F5\_R1

delta-hexatoxin

Aligned protein groups (7,14,16,30) containing a CCC motif and 8 Cys.

37\_1680.3d.comp616\_c0\_seq1\_F1\_R1|spider.q17.comp1850\_c0\_seq1\_F1\_R1

spider.q17.comp6008\_c0\_seq1\_F1\_R1|37\_1680.3d.comp7541\_c0\_seq1\_F2\_R1

spider.q17.comp1543\_c0\_seq2\_F2\_R1|37\_1680.3d.comp423\_c0\_seq1\_F1\_R2|spider.q17.comp1543\_c0\_seq1\_F1\_R2

spider.q17.comp1516\_c0\_seq1\_F6\_R1

## May 9, 2012

Looked for paralog evidence in group 18 using the reads from PAVenom\_spider.q17 . Low number of reads (5-15) observed for major sequence. No evidence for paralogs.

Aligned protein groups containing 10 Cys to look for likely paralogs. They appear to be too divergent

10cys containing groups (3.1, 3.2,4,10,19)

spider.q17.comp1699\_c0\_seq2\_F3\_R3|spider.q17.comp1699\_c0\_seq5\_F3\_R3

spider.q17.comp1699\_c0\_seq3\_F3\_R3

37\_1680.3d.comp452\_c0\_seq2\_F2\_R3|37\_1680.3d.comp452\_c0\_seq1\_F2\_R2|spider.q17.comp518\_c0\_seq1\_F5\_R3|AS646

37\_1680.3d.comp799\_c0\_seq1\_F2\_R4|spider.q17.comp2034\_c0\_seq1\_F5\_R6

AS209

## May 8, 2012

Looked for paralog evidence in group 1 using the reads from PAVenom\_spider.q17 for

Only observed an N to D substitution. This was already supported in the search results by spectra matched by N to n deamidation.

## May 3-6, 2012

Processed the run K121115\_Hi\_dsrak\_ff4f4f4CHEsa\_cccc\_01.raw that is a Lys-c digest of Hi venom after alkylating Cys with 2-methylaziridine. The run has lots of good data. The modified cys bear charge and appear to be cleaved by Lys-C. Some unexplained ions in the cys containing ETD spectra. Especially when pair or triplet of cys nearby.

Otherwise excellent ETD fragmentation

Had to modify msparams\_mill\instrument.txt; pos\_charge RHKNC

Revise SM C++ programs to allow a modification to define a basic residue, instead of resorting to revising instrument.txt.

Modify summaryPP to turn on PIP display

## May 1, 2012

Looked closer at the MS/MS spectra of the longer peptides z = 5,6,7 and noticed that deamidated Asn often in sequence and precursor chi2 values were often poor. Inspection of the MS1 spectra shows that the 12C is often neither being correctly called by the instrument acquisition software nor corrected by SM.

Look into revising SM extractor to calculate chi2 from a merger of the preceeding and following MS1 spectra.

## April 30, 2012

From the summary table it is clear that the proteins which break out into the most isoforms do so with isoforms that are accounted for by a sequence present in Arachnoserver but not in the RNA-Seq data. This may mean that Trinity is over condensing the isoforms, and that there are likely to be more isoforms found amongst the protein groups which are currently defined only by the RNA-Seq data.

David provided the original .fastq files containing the Ion Torrent sequence reads from the venom gland libraries. The datasets tend to have 3-8 million reads. The read lengths are 50-200 base pairs each. I revised the seqdb\extractORFS.pl script to parse and translate them. The Phred quality scores were ignored.

seqdb\ PAVenom\_spider.q17

Using the group 2 omega-hexatoxin Hi2 c-terminal regions and MS-Edman it is clear that several thousand copies of each sequence are present and there are hundreds of copies consistently containing specific isoform AA’s in some positions. Given the short read lengths it seems likely that there would be difficulty assembling full length toxin sequences 40-70 AAs that get individual isoforms that correctly contain several alternate AA’s. the reads do contain some evidence that spans the regions contiaing AA substitutions.

David also provide a copy of an article that helps explain the naming convention for toxins

King et al, Toxicon 52 (2008) 264–276 “A rational nomenclature for naming peptide toxins from spiders

and other venomous animals” .

## April 23-29, 2012

I subjected the data to 1% FDR cutoff and then manually inspected numerous spectra. After organizing the data into proteins there are 24 protein groups, some of which have 2-6 isoforms. All of those groups result in proteins which have 6, 8 or 10 Cys in the mature protein and are <10KDa.

msdataSM\2012venom\spider\ spider\_Hi\_ProteinCentricColumns.xlsx

I created a subset database in SM from the valid accession numbers:

Seqdb\FASTA\_spider\blast2GO\PAVenom\_Arachno\_Hi\_3day.valid.fasta

and used that to run BLAST2GO against the UniProt database.

Seqdb\FASTA\_spider\blast2GO\ blast2go\_20130426\_1323.dat

I exported

Seqdb\FASTA\_spider\blast2GO\spiderTopBlast .xlsx

And used that to make

Seqdb\PAVenom\_Arachno\_Hi\_3day.categories.tsv

To add extra columns in SM reports for:

BLAST2GO-Protein-Name Species-Top-Blast-Hit Top-BLAST-UniProt-ID Top-BLAST-E-Value

numCys Sequence

were also added from the file

Seqdb\FASTA\_spider\blast2GO\PAVenom\_Arachno\_Hi\_3day.valid.xlsx

## April 23, 2012

When examining ETD spectra it became apparent that peaks from neutral loss of the cys side chain were not being assigned properly. I revised the marker ion definition for carabamidomethylation in msparams\_mill\ smconfig.std.xml and repeated the ETD searches.

## December 2012

David gave me several FASTA databases that resulted from several Ion Torrent RNA-Seq datasets that had each been assembled by Brian Haas using Trinity. The 3-day libraries should be enriched in toxins. Not sure why.

Seqdb\FASTA\_spider

We decided to create a concatenated database that contained the protein translations from the following:

Arachnoserver

37\_1680.3d.Trinity.fasta

spider.q17.Trinity.fasta

I started searches of the data in Spectrum Mill using no enyme mode, and allowing for oxidized Met, Gln to pyro-Glu, and deamidation of Asn. The ETD searches allowed for a fragment mass tolerance of 100ppm. In November while looking at Centipede ETD data collected at the same time I noticed that the ETD data mas large mass errors at high mass. THE CID and HCD data are fine. ETD data collected after July 2012 doesn’t have the same problem. In July 2012 the instrument acquisition computer was replaced and a newer rev of the acquisition software installed.

## February 2012

See file venomSummaryMS\_Feb2012.xlsx

Venom sample: funnel web spider, Hadronyche infensa, Queensland Australia

Cysteines: reduced/alkylated with DTT/iodoacetamide

Digestion: none

Filename: K120223\_Hi\_ri\_ff4f4f4\_pccc\_01.raw

LC-MS/MS: Orbitrap Elite with high resolution CID/HCD/ETD triplets on every precursor. Supplemental activation for ETD.

LC: 300A Microsorb 3um C18 KC120222a, DM90 gradient

Despite the lack of digestion, we observed a lot of ragged end peptides. So it looked like the material had a lot of hydrolysis. Those observations were from searches against all sequences in Arachnoserver, which contains a few Hi toxins and accompanying isoforms.