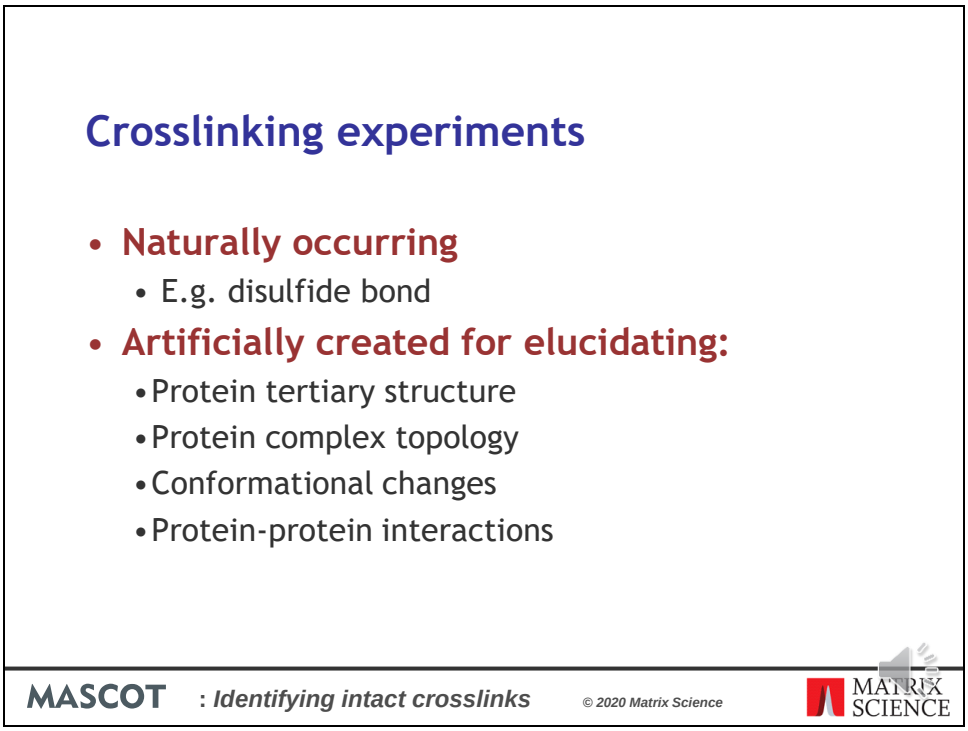
Identifying intact crosslinks with Mascot Server

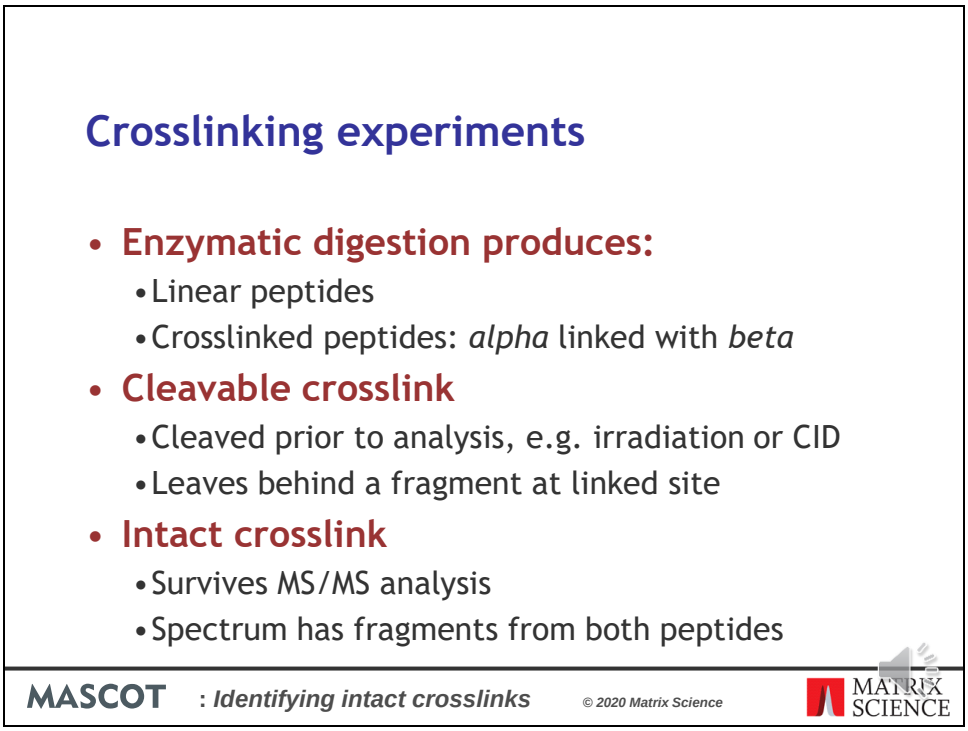
Ville Koskinen
Matrix Science

A new feature in Mascot Server 2.7 is the ability to identify crosslinked peptides from MS/MS data. I'll show you how to use the new feature and discuss validating crosslinked search results.



Crosslinks can be roughly divided into naturally occurring and artificially created. An example of a naturally occurring crosslink is the disulfide bond. In most experiments, disulfide bonds are of no interest, so they are cleaved with a reducing agent prior to analysis. But, if the goal is to study protein folding, identifying the sites linked by a disulfide bond is crucial.

Crosslinks can also be created artificially by a chemical reaction. A crosslink can only be formed between sites that are adjacent in three-dimensional space. The maximum distance depends on the spacer arm length of the linker. So, the locations of the identified links can be used for constraining the three-dimensional shape of the molecule, which has obvious applications in elucidating protein tertiary structure, protein complex topology, conformational changes and protein-protein interactions.



Whether natural or artificial, a typical bottom-up crosslinking experiment starts by digesting the crosslinked proteins. This produces two types of precursor: linear peptides and crosslinked peptides. A linear peptide either doesn't span a linkable site or if it does, the linker did not form a crosslink. Otherwise, the precursor is a crosslinked peptide, where an alpha peptide is linked with a beta peptide. The alpha peptide is conventionally the longer or more massive of the two.

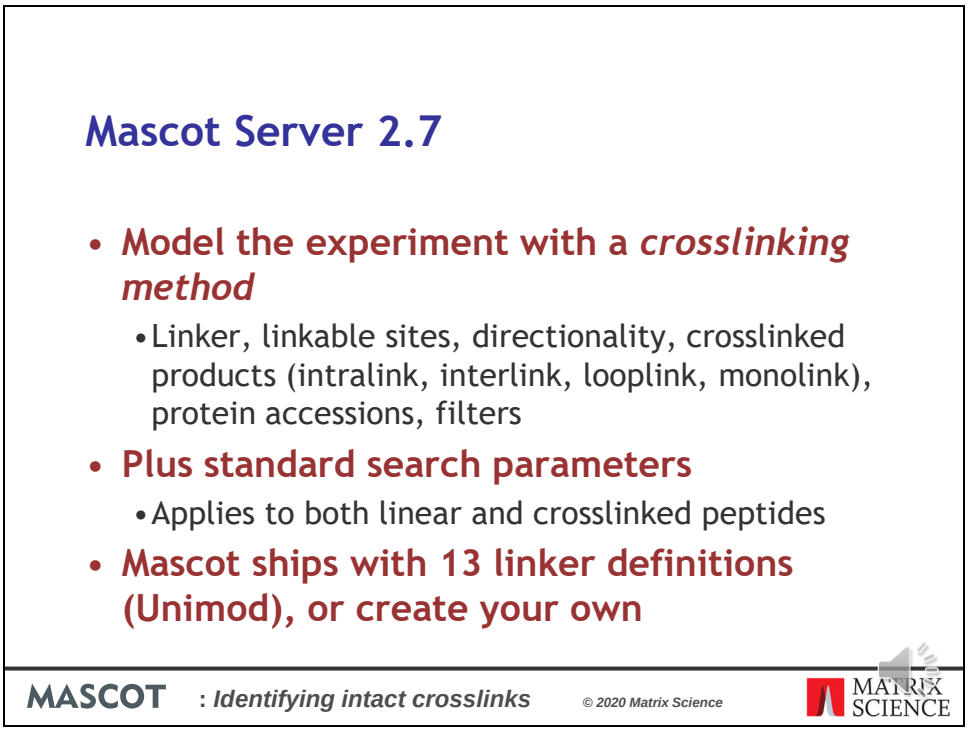
How crosslinked peptides appear to a search engine like Mascot depends on the type of the linker.

Cleavable crosslinks are designed to be cleaved prior to analysis by various means like irradiation or CID. The link cleaves asymmetrically, so that the alpha and beta peptides carry linker fragments with different mass. The fragment is modelled exactly like a variable modification on a linear peptide. The linear peptide matches are then paired up by a postprocessing algorithm, which figures out which peptides come from the same crosslinked precursor.

Conversely, intact crosslinks survive MS/MS analysis. There is additional complexity to the crosslinked spectrum: alpha fragments that contain the linked site have a mass offset equal to the beta peptide mass plus the linker molecule. Vice versa for beta

fragments. The search engine must be able to fragment crosslinked peptides *in silico*. However, the advantage is, no postprocessing step is needed.

The boundary between cleavable and intact is a bit fuzzy. If the cleaving efficiency isn’t 100%, there will be some precursors with an intact link. Similarly, an intact linker can fail to form a link, which leaves behind a linker fragment.



Because of the great variety and complexity in crosslinking experiments, we've chosen an approach similar to quantitation. We want to keep the software universal and configurable, and not tied to a specific experimental workflow.

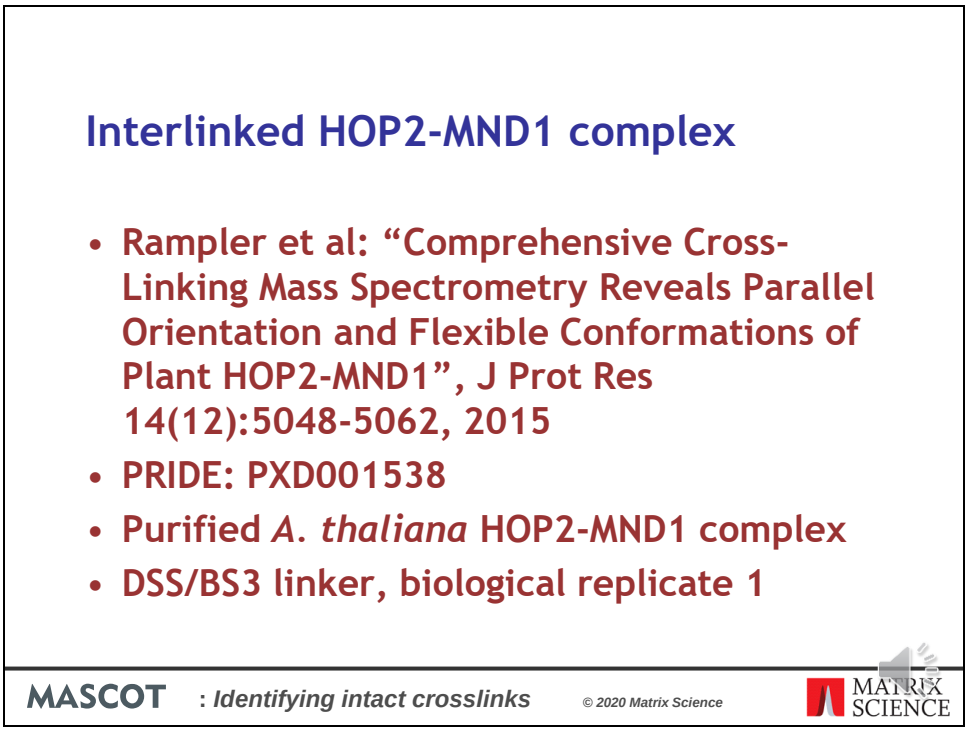
The crosslinking experiment is modeled with a crosslinking method. The method defines which linker is used, which linkable sites are considered, linker directionality, what are the expected crosslinked products and which protein or proteins are being linked. There are also a few filters available.

There are four possible crosslinked products. An intralink is an intact crosslinked peptide within the same protein, while an interlink is an intact link between two different proteins. A looplink is an intact link within a peptide. Finally, monolinks model cleavable linker fragments as well as dead-end or quenched links where only one end is attached to a peptide. I'll show examples of crosslinked products later.

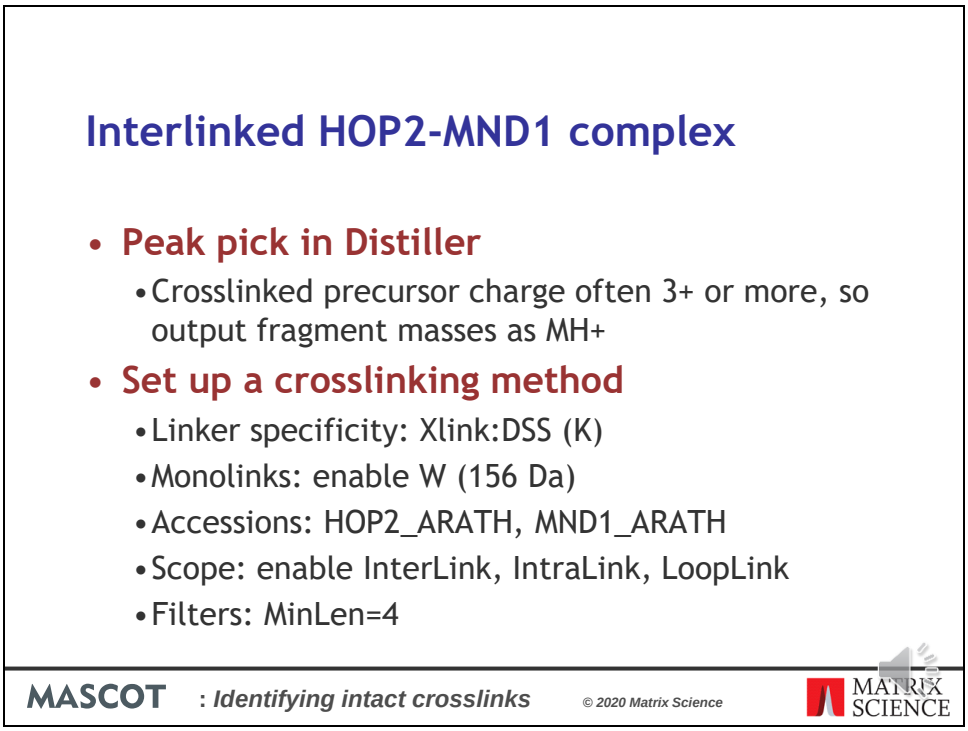
The standard search parameters like enzyme, fixed mods and variable mods are applied to both linear peptides and crosslinked peptides.

The linker definitions are standard Unimod entries. Mascot 2.7 ships with 13

definitions: 7 for intact crosslinks and 6 for cleavable crosslinks. You can easily create your own definition for any linker chemistry.



I'll illustrate intact crosslinking with this data set from the Mechtler lab. The data is available as the PRIDE project PXD001538. The study uses a purified HOP2-MND1 complex from *Arabidopsis thaliana*, which is crosslinked using three different linkers. I've chosen the biological replicate 1 linked with disuccinimidyl suberate (DSS). DSS has two identical reactive groups, so it's homobifunctional. It's amine reactive and creates intact crosslinks between lysines and protein N-term. The study was designed so that there are only lysine-lysine links.

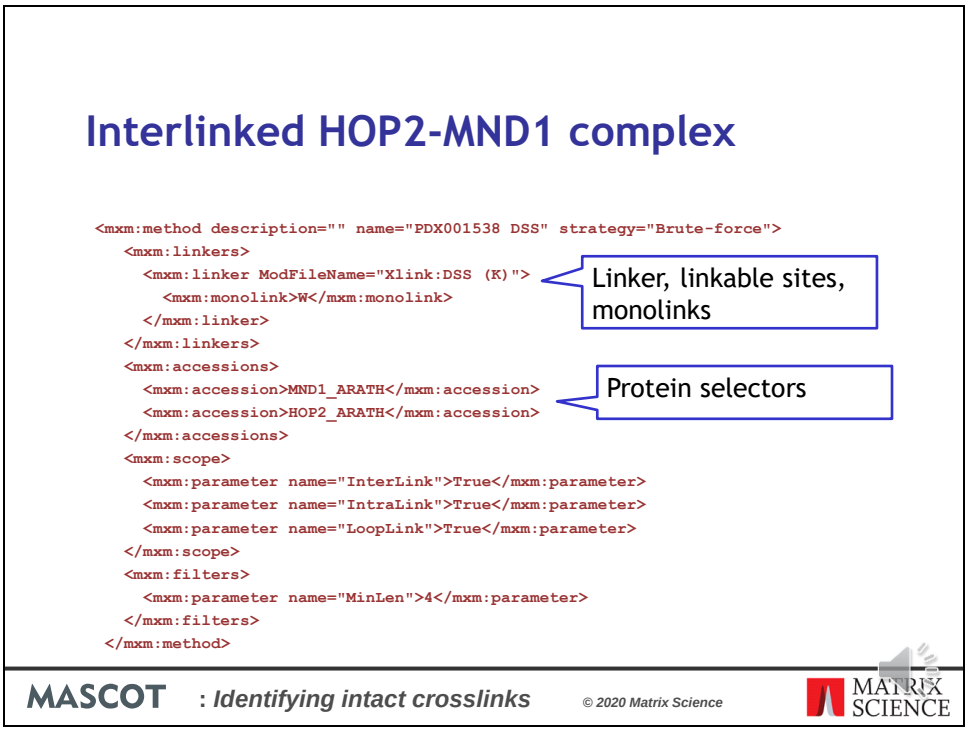


The raw data was peak picked in Mascot Distiller. Crosslinked precursors often have charge of 3+ or more, which means the product ions could also be highly charged. It's best to decharge the fragments by choosing to output fragment masses as MH+.

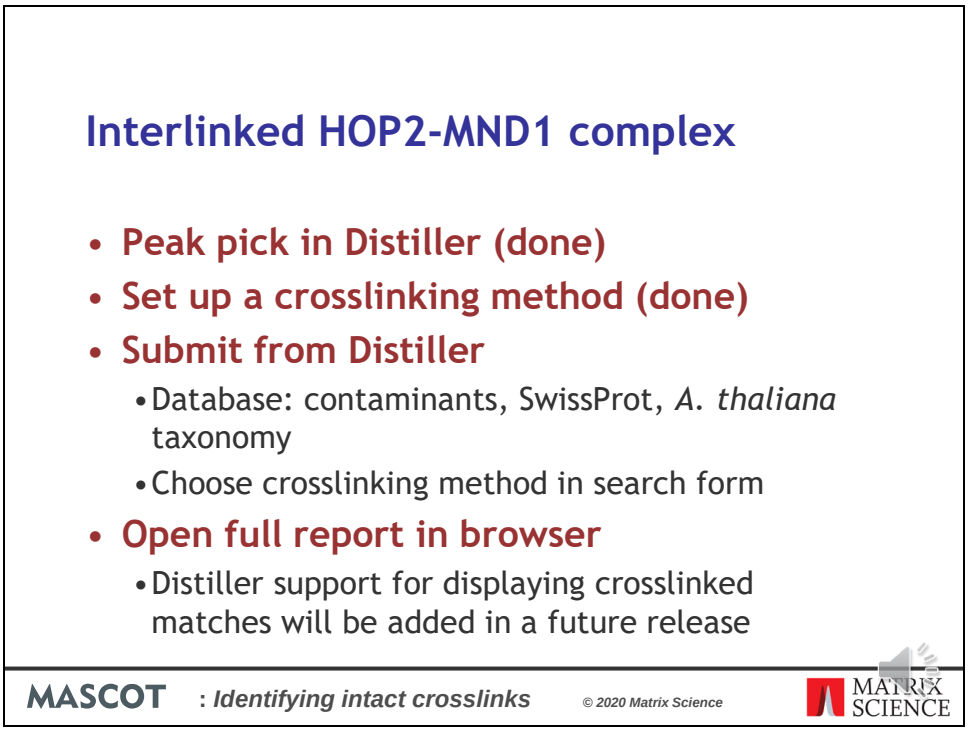
Setting up the crosslinking method is straightforward. DSS has a Unimod definition titled Xlink:DSS, and we only want the (K) specificity for lysine-lysine links. The authors determined only one of the three monolinks is possible, in this case the water quenched 156 Da monolink. This has the code W in the Unimod definition.

The accessions in the crosslinking method act as selectors and instruct Mascot to choose only the listed proteins for crosslinking. The sequences are in SwissProt, HOP2_ARATH and MND1_ARATH, so we can just list them here and select SwissProt as the search database when the search is submitted.

The crosslinked products to consider are protein interlinking and intralinking, and peptide looplinking. Finally, allow the beta peptide to be as short as 4 residues. If no MinLen is specified, Mascot filters out alpha and beta peptides shorter than MinPepLenInSearch, which is 7 by default.

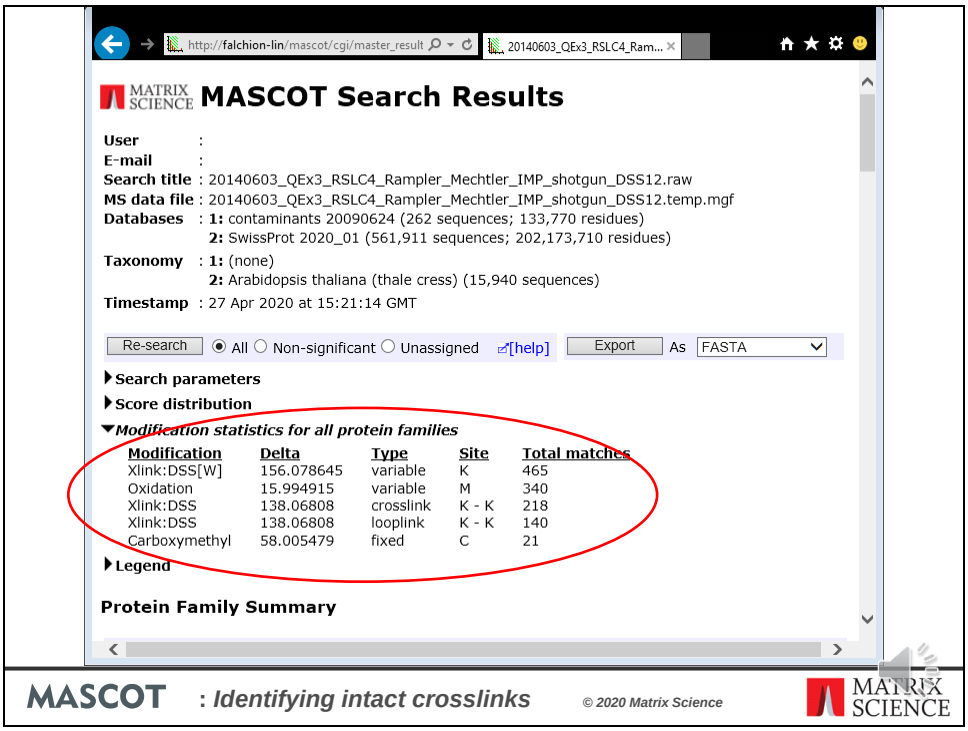


The crosslinking method is a simple XML file, shown here. Monolinks are toggled in the mxm:linker section. The mxm:accession elements can optionally specify a database name. If there is no database name, Mascot chooses proteins for crosslinking by accession, regardless of database. The scope and filters sections contain simple key-value parameters.

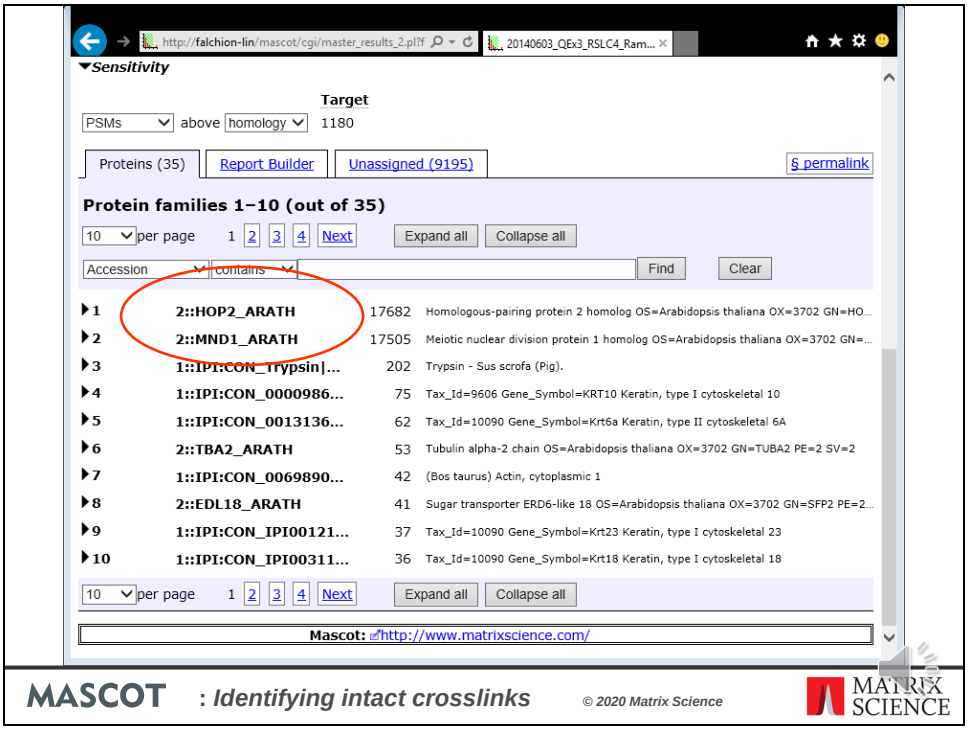


So, the raw data has been peak picked and we have a crosslinking method. The search is submitted as normal from Distiller. Since the two proteins are in SwissProt, I've chosen to search SwissProt with *A. thaliana* taxonomy for convenience. Also, remember to include a contaminants database. Choose the right crosslinking method in the search form.

Once the search finishes, choose to open the full report in browser. At the moment, Distiller can only display linear peptides. Support for crosslinked matches will be added in a future release.

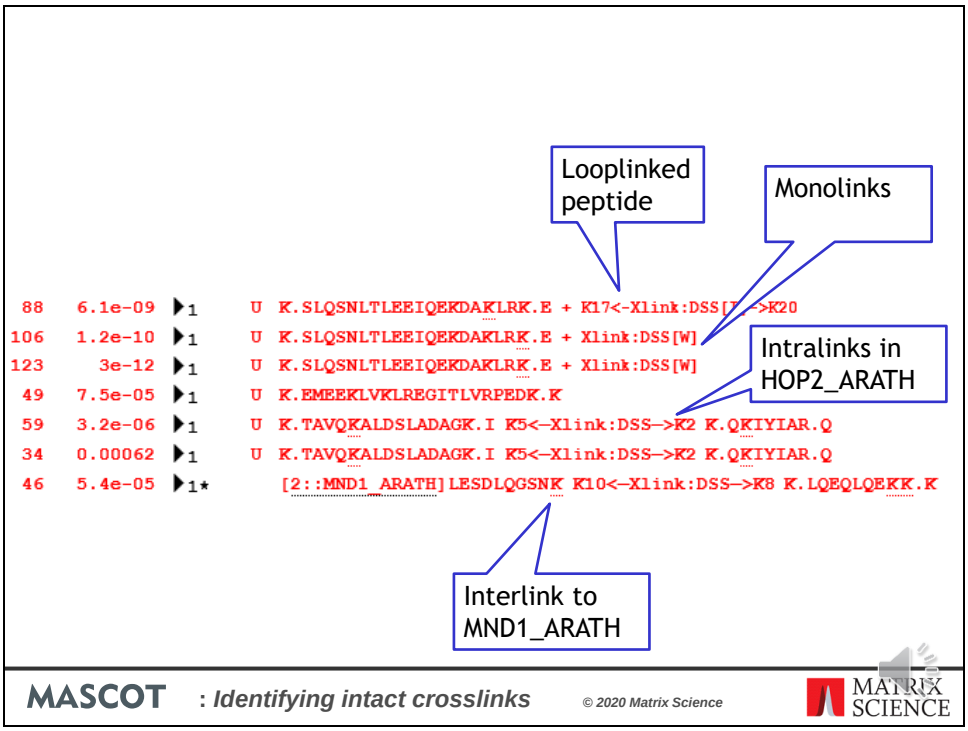


The Protein Family Summary report has been extended to display crosslinked search results. The modification statistics table gives counts of discovered intact links, monolinks and looplinks. Xlink:DSS[W] on the first row is the 156Da monolink enabled in the crosslinking method. The crosslink and looplink rows display the intact linker mass as well as the linked sites.



Protein inference is not affected by interlinking. Because HOP2_ARATH and MND1_ARATH don't share any significant peptide matches, protein inference puts them in different families, regardless of whether there are any interlinked peptide matches.

Notice also the small amounts of evidence for other Arabidopsis proteins in the sample, despite the high levels of purification.



Expand the first protein hit, HOP2_ARATH. The peptide match table displays crosslinking data in a compact form, and I've selected a few example rows here.

On the first line is a looplinked peptide. A looplink is an intact link between two residues in the same peptide. Here the intact link is between lysines K17 and K20. Looplinks use the symbol [I] in brackets to indicate it's an intact link.

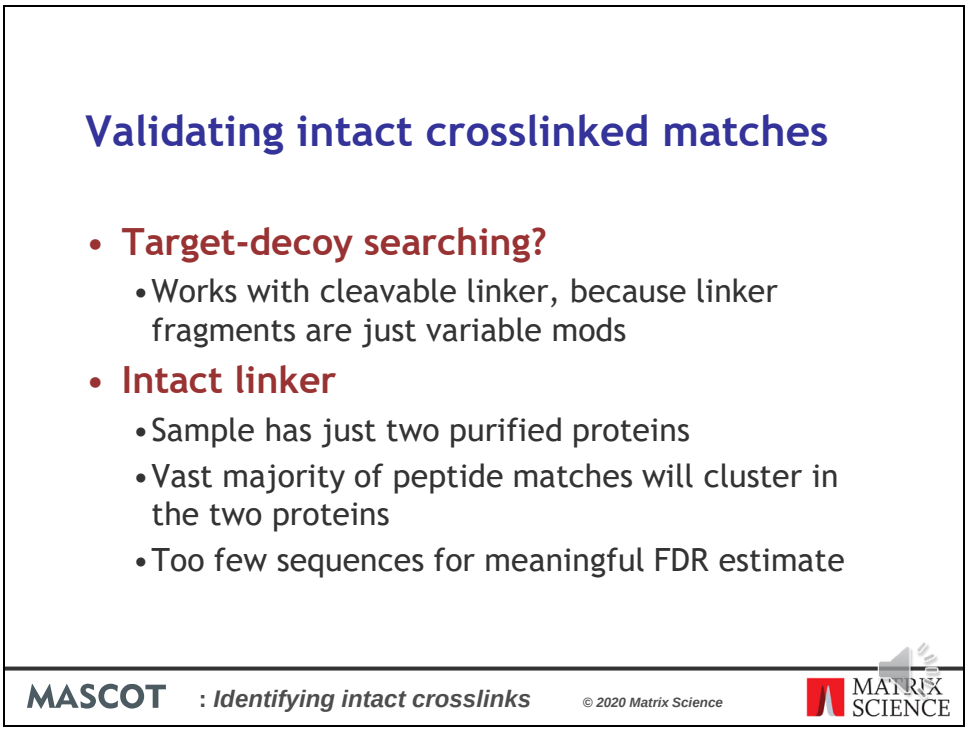
The next two rows show matches to the same peptide sequence, this time with the W monolink. Monolinks are variable modifications generated from the linker definition. With cleavable linkers, this is usually the only crosslinked product of interest, since monolinks represent the linker fragments. Monolinks can also be used for modelling dead-end or quenched links, such as the water quenched monolink here.

Further down, there are three matches with an intact crosslink. The first two are protein intralinks, which means it's an intact link between two non-overlapping peptides within a single protein.

The last line shows a match to a protein interlink. An interlink is an intact link between two peptides from different proteins. The alpha peptide is in MND1_ARATH

and the beta peptide in HOP2_ARATH.

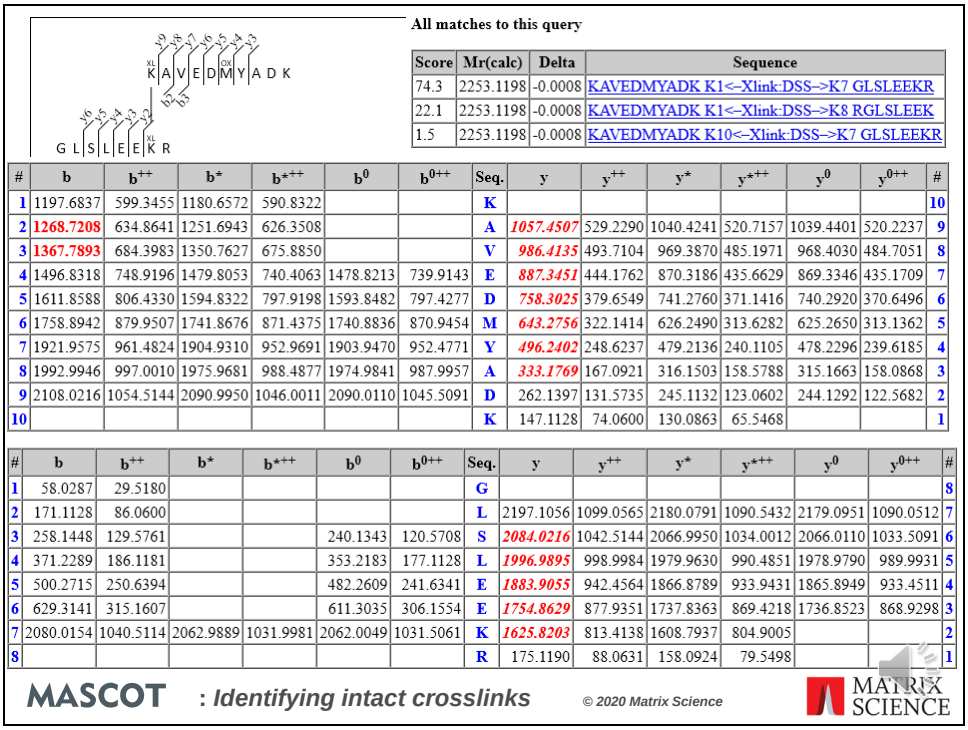
The scores for all the matches are standard Mascot scores, and the identity and homology thresholds are calculated using the same statistical model as for linear peptide matches.



Which of the matches on the previous slide are correct? The usual approach is to do a target-decoy search, which allows you to estimate the false discovery rate. This approach works if you're using a cleavable linker. The linker fragment is just a variable modification and all the peptide matches are ordinary linear peptides., so standard FDR estimation techniques work well in this case.

However, when using intact linkers, the search space is often too small for a meaningful estimate of the FDR. In the present case, the sample is expected to contain only the two purified proteins. If you set up a two-sequence database and search ten thousand MS/MS spectra against it, no statistical model can give a meaningful estimate of the FDR. In this case, I've used as the target database the fifteen thousand *A. thaliana* proteins in SwissProt, plus a couple hundred contaminant sequences, which allows incorrect matches to distribute a bit more evenly. Even so, the peptide matches completely saturate the available peptide sequences from the two target proteins, and there are few matches in other proteins.

The Mascot statistical model is robust and will give you reasonable separation between correct and incorrect matches, even in these adverse conditions. Nonetheless, it's prudent to keep a sceptical mind. Let's look at a few example spectra.



All matches to this query

Score	Mr(calc)	Delta	Sequence
74.3	2253.1198	-0.0008	KAVEDMYADK K1<–Xlink:DSS–>K7 GLSLEEKR
22.1	2253.1198	-0.0008	KAVEDMYADK K1<–Xlink:DSS–>K8 RGLSLEEK
1.5	2253.1198	-0.0008	KAVEDMYADK K10<–Xlink:DSS–>K7 GLSLEEKR

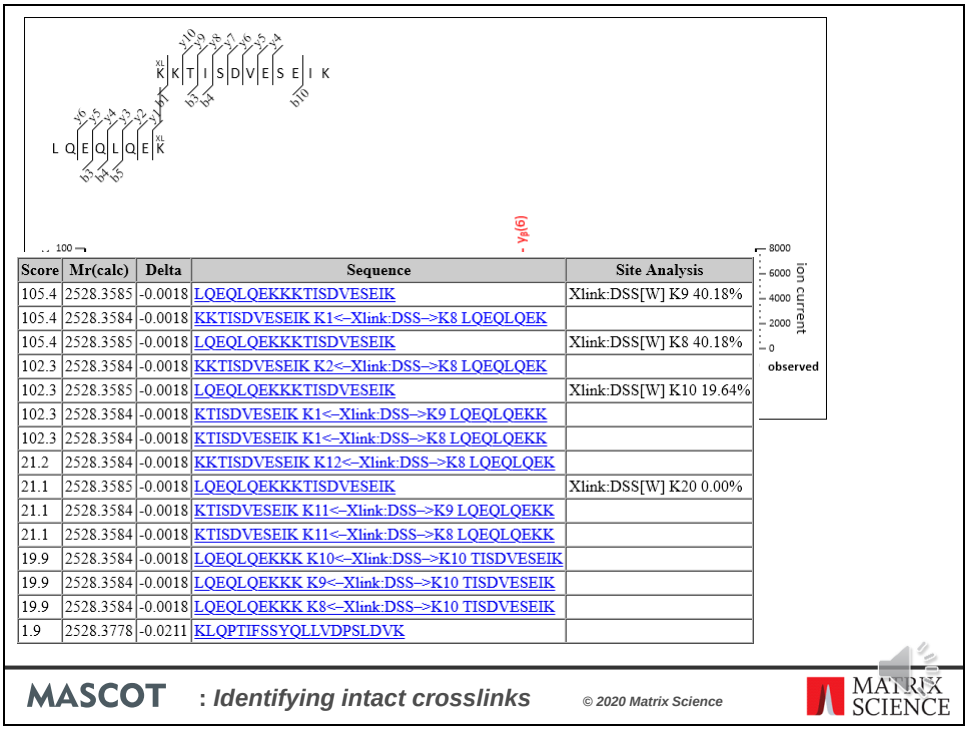
#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	1197.6837	599.3455	1180.6572	590.8322			K							10
2	**1268.7208**	634.8641	1251.6943	626.3508			A	***1057.4507***	529.2290	1040.4241	520.7157	1039.4401	520.2237	9
3	**1367.7893**	684.3983	1350.7627	675.8850			V	***986.4135***	493.7104	969.3870	485.1971	968.4030	484.7051	8
4	1496.8318	748.9196	1479.8053	740.4063	1478.8213	739.9143	E	***887.3451***	444.1762	870.3186	435.6629	869.3346	435.1709	7
5	1611.8588	806.4330	1594.8322	797.9198	1593.8482	797.4277	D	***758.3025***	379.6549	741.2760	371.1416	740.2920	370.6496	6
6	1758.8942	879.9507	1741.8676	871.4375	1740.8836	870.9454	M	***643.2756***	322.1414	626.2490	313.6282	625.2650	313.1362	5
7	1921.9575	961.4824	1904.9310	952.9691	1903.9470	952.4771	Y	***496.2402***	248.6237	479.2136	240.1105	478.2296	239.6185	4
8	1992.9946	997.0010	1975.9681	988.4877	1974.9841	987.9957	A	***333.1769***	167.0921	316.1503	158.5788	315.1663	158.0868	3
9	2108.0216	1054.5144	2090.9950	1046.0011	2090.0110	1045.5091	D	262.1397	131.5735	245.1132	123.0602	244.1292	122.5682	2
10							K	147.1128	74.0600	130.0863	65.5468			1

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	58.0287	29.5180					G							8
2	171.1128	86.0600					L	2197.1056	1099.0565	2180.0791	1090.5432	2179.0951	1090.0512	7
3	258.1448	129.5761			240.1343	120.5708	S	***2084.0216***	1042.5144	2066.9950	1034.0012	2066.0110	1033.5091	6
4	371.2289	186.1181			353.2183	177.1128	L	***1996.9895***	998.9984	1979.9630	990.4851	1978.9790	989.9931	5
5	500.2715	250.6394			482.2609	241.6341	E	***1883.9055***	942.4564	1866.8789	933.9431	1865.8949	933.4511	4
6	629.3141	315.1607			611.3035	306.1554	E	***1754.8629***	877.9351	1737.8363	869.4218	1736.8523	868.9298	3
7	2080.0154	1040.5114	2062.9889	1031.9981	2062.0049	1031.5061	K	***1625.8203***	813.4138	1608.7937	804.9005			2
8							R	175.1190	88.0631	158.0924	79.5498			1

MASCOT : *Identifying intact crosslinks* MATRIX SCIENCE

Here's a Peptide View report of a crosslinked match with excellent fragmentation from both alpha and beta, with score 74. Further down the page, there is a fragmentation table for the alpha peptide and another for the beta peptide. The fragment masses have no overlap between the two, which shows both peptides have independent evidence.

At the bottom of the page, you can see the query also matches a slightly different beta peptide. However, the rank 2 score is much lower, only 22, and there's no ambiguity in the alpha peptide, so we can be confident the top match is the correct match.

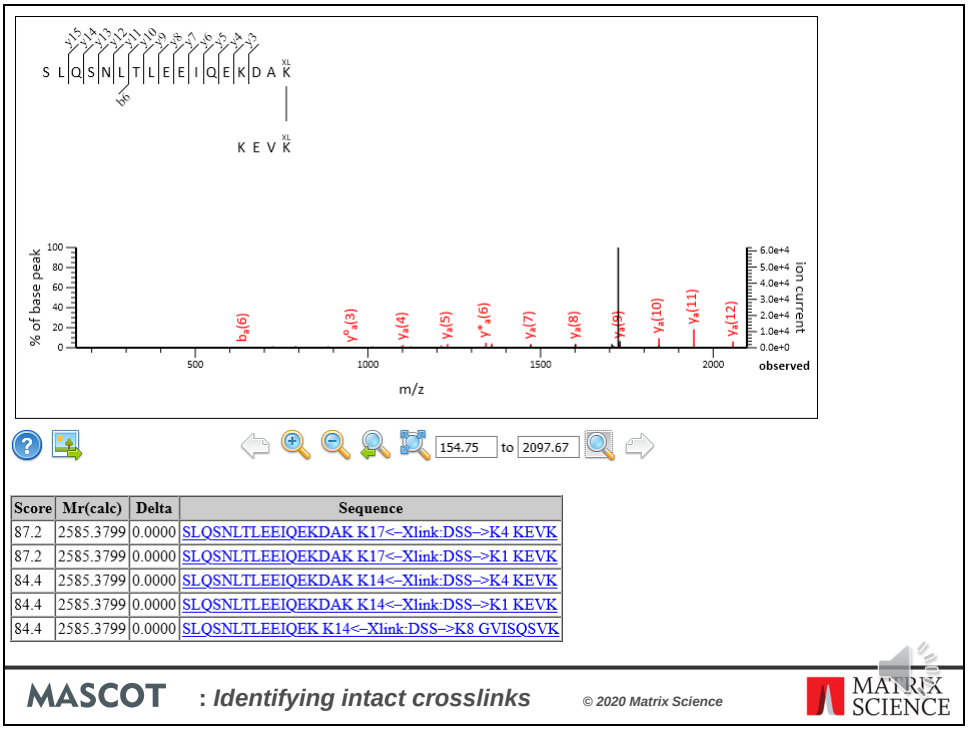


Score	Mr(calc)	Delta	Sequence	Site Analysis
105.4	2528.3585	-0.0018	LQEQLQEKKKTISDVESEIK	Xlink:DSS[W] K9 40.18%
105.4	2528.3584	-0.0018	KKTISDVESEIK K1<–Xlink:DSS–>K8 LQEQLQEK	
105.4	2528.3585	-0.0018	LQEQLQEKKKTISDVESEIK	Xlink:DSS[W] K8 40.18%
102.3	2528.3584	-0.0018	KKTISDVESEIK K2<–Xlink:DSS–>K8 LQEQLQEK	
102.3	2528.3585	-0.0018	LQEQLQEKKKTISDVESEIK	Xlink:DSS[W] K10 19.64%
102.3	2528.3584	-0.0018	KTISDVESEIK K1<–Xlink:DSS–>K9 LQEQLQEKK	
102.3	2528.3584	-0.0018	KTISDVESEIK K1<–Xlink:DSS–>K8 LQEQLQEKK	
21.2	2528.3584	-0.0018	KKTISDVESEIK K12<–Xlink:DSS–>K8 LQEQLQEK	
21.1	2528.3585	-0.0018	LQEQLQEKKKTISDVESEIK	Xlink:DSS[W] K20 0.00%
21.1	2528.3584	-0.0018	KTISDVESEIK K11<–Xlink:DSS–>K9 LQEQLQEKK	
21.1	2528.3584	-0.0018	KTISDVESEIK K11<–Xlink:DSS–>K8 LQEQLQEKK	
19.9	2528.3584	-0.0018	LQEQLQEKKK K10<–Xlink:DSS–>K10 TISDVESEIK	
19.9	2528.3584	-0.0018	LQEQLQEKKK K9<–Xlink:DSS–>K10 TISDVESEIK	
19.9	2528.3584	-0.0018	LQEQLQEKKK K8<–Xlink:DSS–>K10 TISDVESEIK	
1.9	2528.3778	-0.0211	KLQPTIFSSYQLLVDPSLDVK	

Here is another crosslinked peptide match with a high score and pretty nice fragmentation. At such a high score, 105, it looks very promising, but I have to dash your hopes.

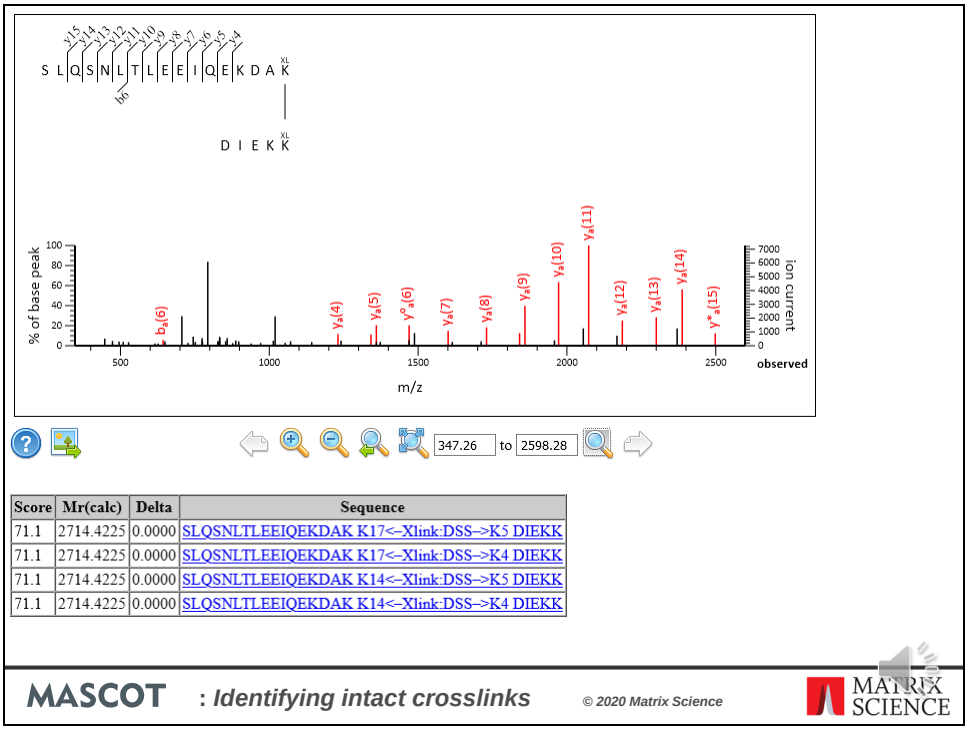
Here are the matches at the other ranks. It turns out, beta and alpha are sequential in the protein sequence. The concatenated beta-alpha sequence with a water quenched monolink has exactly the same score and equally convincing fragment matches. The location of the potential monolink or crosslink is also in question. The MS/MS spectrum has no evidence to differentiate between the alternatives, so choosing the correct match requires consulting structural data.

This also shows the importance of presenting an integrated report. If the search engine matched only crosslinked or only linear peptides, the ambiguity could be easily missed.



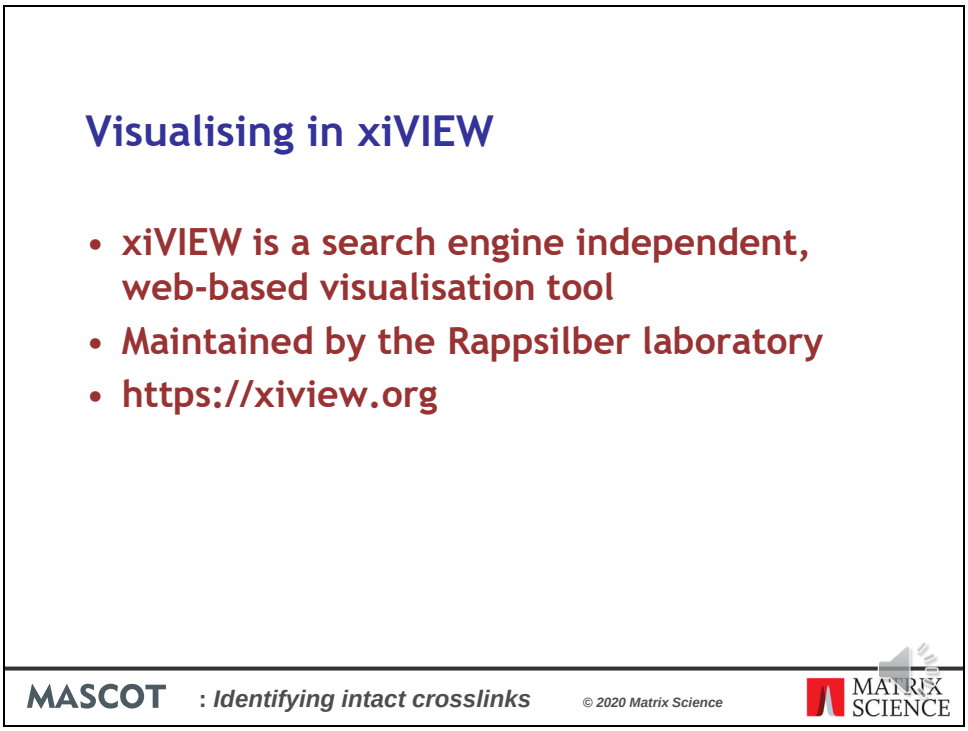
Score	Mr(calc)	Delta	Sequence
87.2	2585.3799	0.0000	SLQSNLTLEEIQEKDAK K17<--Xlink:DSS-->K4 KEVK
87.2	2585.3799	0.0000	SLQSNLTLEEIQEKDAK K17<--Xlink:DSS-->K1 KEVK
84.4	2585.3799	0.0000	SLQSNLTLEEIQEKDAK K14<--Xlink:DSS-->K4 KEVK
84.4	2585.3799	0.0000	SLQSNLTLEEIQEKDAK K14<--Xlink:DSS-->K1 KEVK
84.4	2585.3799	0.0000	SLQSNLTLEEIQEK K14<--Xlink:DSS-->K8 GVISQSVK

Here's a crosslinked match where the identity of the beta peptide should be questioned. The alpha peptide has a nice sequence ladder, but there's no fragmentation from the beta. There's an almost as good a match at rank 5, where the beta peptide is completely different. The alternatives score almost the same because there are no alpha fragments that span the linker site. The two alpha peptides match exactly the same observed fragments.



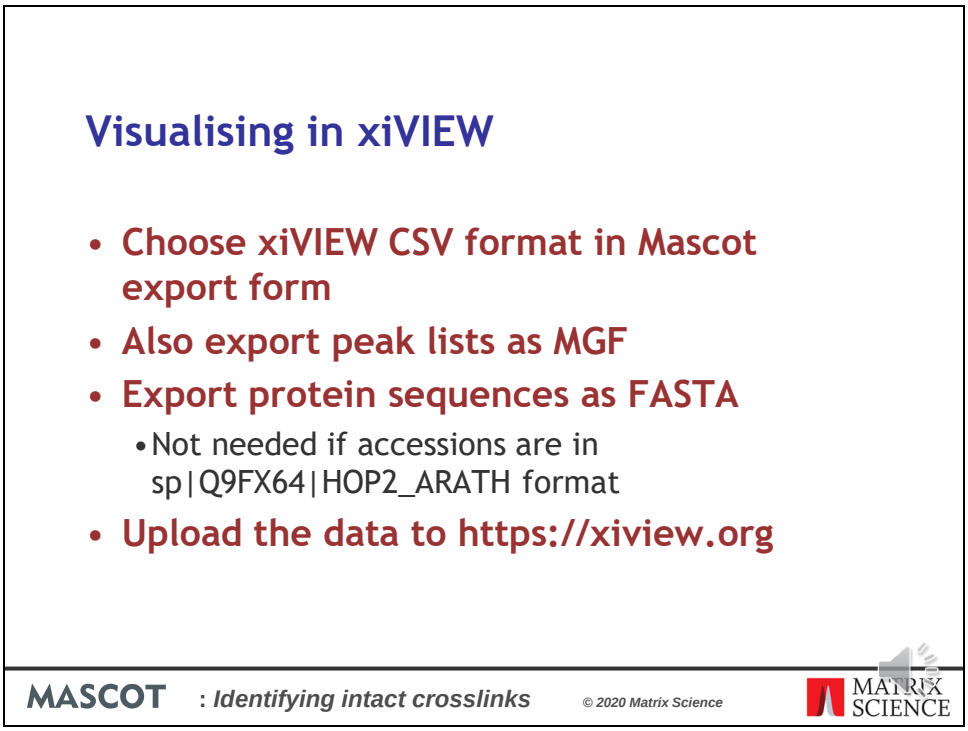
Score	Mr(calc)	Delta	Sequence
71.1	2714.4225	0.0000	SLQSNLTLEEIQEKDAK K17<–Xlink:DSS–>K5 DIEKK
71.1	2714.4225	0.0000	SLQSNLTLEEIQEKDAK K17<–Xlink:DSS–>K4 DIEKK
71.1	2714.4225	0.0000	SLQSNLTLEEIQEKDAK K14<–Xlink:DSS–>K5 DIEKK
71.1	2714.4225	0.0000	SLQSNLTLEEIQEKDAK K14<–Xlink:DSS–>K4 DIEKK

Finally, this crosslinked match is very similar to the previous one with no fragmentation from the beta peptide. But in this case, the search space has no other alternative for beta. If we assume the search space is complete, there's no reason to doubt the identity of the beta peptide, even if there are no beta fragments in the spectrum.



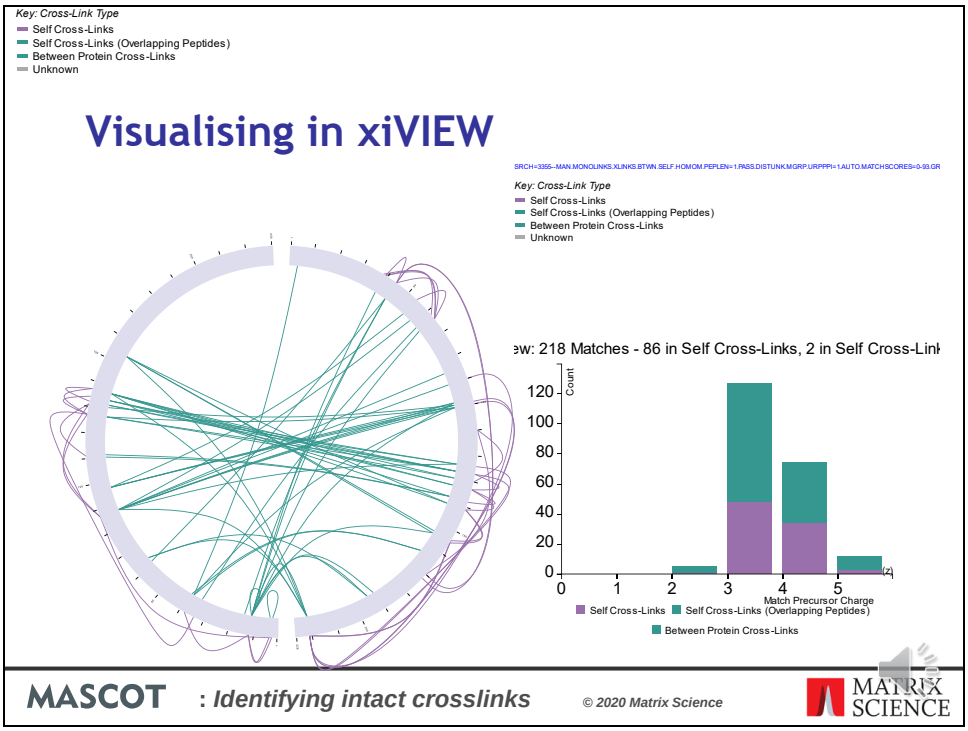
Other common ways to validate intact crosslinked matches is by comparing with existing annotations, and comparing with structural data. If the proposed crosslink would have a length longer than the linker's spacer arms, it's either a false positive or the structural data is incorrect. Mascot doesn't currently make use of structural data, so the validation must be done with external tools.

One way to visualise the results is by viewing them in xiVIEW. This is a search engine independent tool maintained by the Rappsilber laboratory.



Export the results in xiVIEW CSV format from the Mascot export form. If you want to view the annotated spectra in xiVIEW, also export the peak lists as MGF. Finally, export the sequences as FASTA. xiVIEW can fetch the protein sequences and structural data automatically. Unfortunately, the default accession format for SwissProt in Mascot is the AC format, and xiVIEW expects the raw accession.

Upload the files to xiview.org.



Here's an example circular plot of the crosslinks between the two proteins, based on the Mascot search data. xiVIEW can also plot histograms like this one showing the crosslinked match charge states. As expected, nearly all of the crosslinked matches have charge at least 3+ or higher.

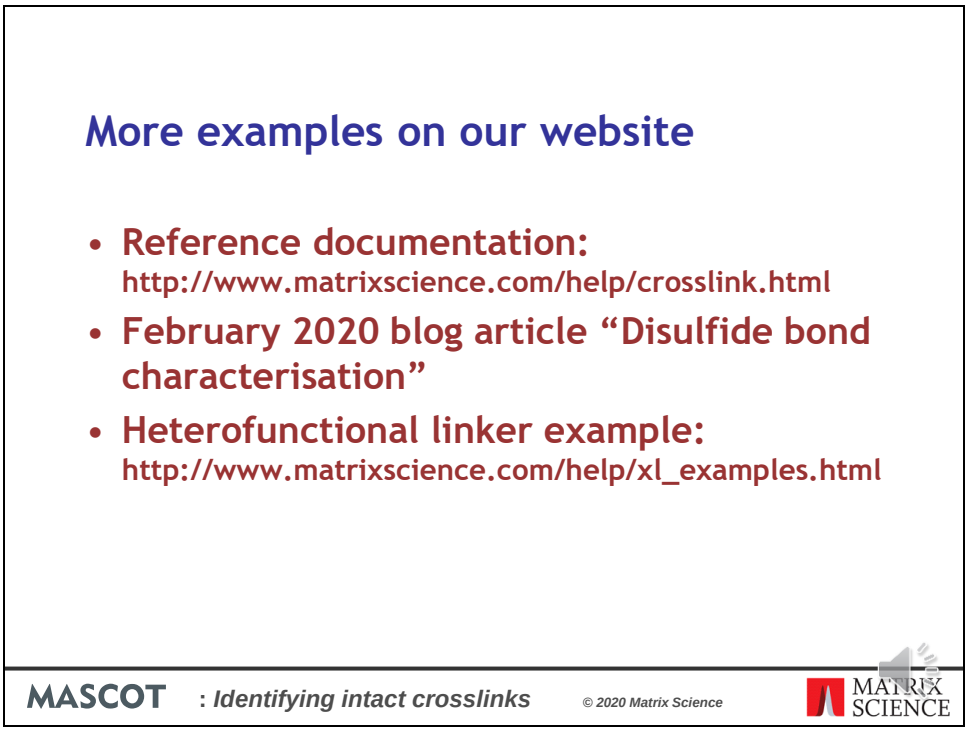
Unfortunately, there is currently no crystallography data for the HOP2-MND1 complex from *Arabidopsis thaliana*, so I can't show you the 3D view.

Visualising in xiVIEW

	A	B	C	D	E	F	G	H	I	J	K
1	Id	Protein1	PepPos1	PepSeq1	LinkPos1	Protein2	PepPos2	PepSeq2	LinkPos2	Score	CrossLinkerModMass
2	q1645_p1_1	MND1_ARATH	6	GLSLEEKR	7	HOP2_ARATH	189	DVKELK	6	16.18	138.06808
3	q2520_p1_1	MND1_ARATH	6	GLSLEEK	7	MND1_ARATH	35	M15.994915GPK	4	15.98	138.06808
4	q3722_p1_1	MND1_ARATH	6	GLSLEEK	7	MND1_ARATH	2	SKKR	3	16.33	138.06808
5	q3723_p1_1	MND1_ARATH	6	GLSLEEK	7	MND1_ARATH	2	SKKR	3	16.07	138.06808
6	q4017_p1_1	MND1_ARATH	4	KRGLSLEEK156.078645R	1	MND1_ARATH	1	M15.994915SKK156.078645	3	24.69	138.06808
7	q4974_p1_1	MND1_ARATH	120	TEALTQLK	8	MND1_ARATH	31	ELEK	4	33.83	138.06808
8	q5157_p1_1	MND1_ARATH	6	GLSLEEK	7	HOP2_ARATH	39	KTAVQK	6	20.99	138.06808
9	q5158_p1_1	MND1_ARATH	6	GLSLEEK	7	HOP2_ARATH	39	KTAVQK	6	28.67	138.06808
10	q5284_p1_1	MND1_ARATH	6	GLSLEEKR	7	HOP2_ARATH	221	KRPR	1	24.18	138.06808
11	q5371_p1_1	MND1_ARATH	6	GLSLEEK	7	HOP2_ARATH	189	DVKELK	6	34.83	138.06808
12	q6018_p1_1	MND1_ARATH	6	GLSLEEKR	7	HOP2_ARATH	39	KTAVQK	6	33.64	138.06808

The xiVIEW CSV format is a simple tabular format, which you can load in Excel or LibreOffice. Modifications like oxidation are encoded as a delta mass within the peptide sequence. The CSV file can be edited before uploading to xiVIEW, for example if you want to remove a putative crosslink because fails a quality check.

xiVIEW can also read crosslink data in mzIdentML 1.2 format, which we will support in a future release of Mascot.



If you'd like to see more examples of crosslinked searches, we have some on our website.