MS Annika User Manual

Micha Johannes Birklbauer, Georg Pirklbauer, Daniela Borgmann, Viktoria Dorfer, Stephan Winkler

MS Annika is a crosslink search engine based on MS Amanda, aimed at identifying crosslinks of MS2-cleavable crosslinkers from MS2 and MS3 spectra.

Nodes

MS Annika consists of four nodes, one of which is optional:

- The **MS Annika Detector** is responsible for identifying spectra that are likely to contain crosslinks.
- The **MS Annika Search** is the actual database search engine.
- The **MS Annika Validator** calculates and applies CSM- and Crosslink-level false discovery rates and persists relevant CSMs.
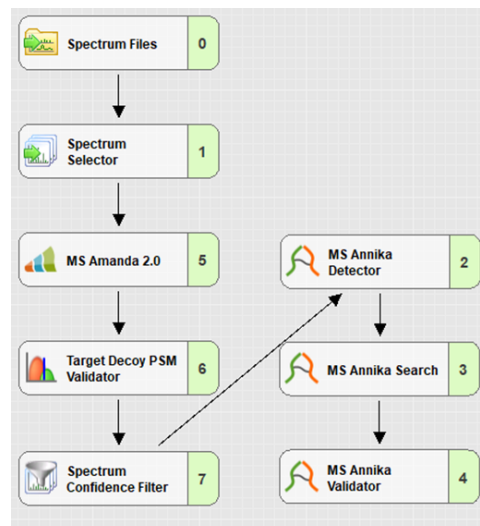- The **MS Annika XiView Exporter** can be used to export crosslinks to the xiView-specific format [1].

In the following, we describe the functionalities of these nodes as well as their parameters.

Figure 1: A typical MS Annika workflow. MS Amanda is not required but will identify linear peptides from spectra with no cross-link evidence.

MS Annika Detector

The MS Annika Detector identifies cross-link spectra. This node splits input spectra into spectra that are likely to contain crosslinks and ones that are not. Spectra classified as unlikely to contain crosslinks can be searched in a normal search engine by attaching one to the Detector node (for example MS Amanda [2]). All other spectra are subjected to crosslink search. Parameters for identifying crosslinks can be adjusted in the Detector node:

- *MS2 tolerance:* Tolerance used to identify if a pair of peaks (a *doublet*) is in range to be a crosslink doublet. Allowed doublet distances are calculated as the difference between heavy and light part, which can be set in the crosslink modification in the chemical modifications and the losses described in *Crosslink Modification Additions*. This tolerance is for identifying and validating Doublets.
- *Minimum Charge:* The minimum charge a peak must have to be considered in the identification of doublets. When setting the value to 0, all peaks will be used, even if the instrument or raw file reader attributed a 0 charge to the ion. We assume that these peaks carry a single charge.
- *Use monoisotopic mass:* Defines whether the monoisotopic or average amino acid mass should be used. Used at different stages in the search.
- *Precursor mass offset:* Charges between 0 and n (the value set here) are considered for the doublet search. For example: If set to 2 the MS1 precursor, the MS1 precursor minus one proton mass, and the MS1 precursor minus two proton masses are considered when deciding if a combination of doublets is a valid doublet.
- *Use theoretic MS1 peaks:* If set to false, only peaks found in the MS1 spectrum are used in the precursor mass offset calculation. If set to true, the precursor mass is used to calculate theoretical peaks with up to n isotope shifts, even if the peak is not in the MS1 spectrum.
- *Crosslink Modification:* The crosslink modification is created using the Chemical Modifications interface (*Administration > Chemical Modifications*). Create a chemical modification and fill out all the required fields. Then, mark the modification and click the *Extended Properties* button at the top of the list. Create the two fragments of the crosslinker. It is also crucial to provide the connected fragments via the *Extended Properties*

window. **If more than one connection is provided, the program will notify the user about the used connection (first connection in the table). It's also important that the first connection is the one with the shortest doublet distance e.g., Alkene – Thiol instead of Alkene – Sulfenic Acid for DSSO.**

- *N-Terminal Crosslink Modification:* Set this value to true if you want to search for protein N-terminal crosslinking sites.
- *Additional Crosslink Doublet Distances:* Additional distances to be considered between the light and the heavy part of the linker. Use this if the linker can have additional losses, such as a water loss as for example exhibited by DSSO. The default doublet distance for DSSO is the distance between the alkene and the thiol fragment at 31.9721 Da – if you want to additionally consider doublets originating from the alkene and sulfenic acid fragments you would need to add their distance here, which is 49.982635 Da. Possible values can be set in the configuration page for the MS Annika Detector (*Administration > Configuration > MS Annika Detector > Additional Crosslink Doublet Distances*). Each line is a possible value. The values are parsed when the Detector node is run and will result in a warning if they cannot be parsed. Attention: If an empty line is included in the possible values, Proteome Discoverer shows an error message! In this case, check the settings (*Administration > Configuration > MS Annika Detector > Additional Crosslink Doublet Distances*) and make sure no empty lines are included.
- *Diagnostic Ions:* The *m/z* value for ions can be set here. These ions are often found in cross-link spectra and are different for each linker. The presence of diagnostic ions in a spectrum increases the likeliness for that spectrum to be selected for the cross-link search. MS Annika also reads diagnostic ions directly from the crosslinker modification definition.
- *Doublet Pair Selection:* One of evidence mode, indication mode or combined mode. Determines how cross-link spectra are searched for cross-link information.
- *top N most intense doublets:* Since indication mode identifies a substantial number of doublets, this parameter allows for the selection of only the most intense ones.
- *Try infer missing charge states:* If this option is enabled (by default) any peak in any given spectrum that does not have an associated charge state will have its charge calculated by its isotope distribution. This is especially important when dealing with MGF files, as MGF does not store fragment ion charge states. Disabling this parameter will increase search speed but MS Annika will not be able to deal with MGF data then.
- *Maximum considered fragment ion charge:* The maximum charge considered for calculating the associated charges of peaks without annotated charge states (used in the algorithm for the above parameter).
- *Persist Doublets:* A Boolean value determining whether found Doublets should be persisted. If set to true, the found doublets will be reported as a result table in Proteome Discoverer. This is false by default, since it is possible that MS Annika identifies a lot of doublets, resulting in huge result files.

Additional parameters in the Configuration Tab:

- *Additional Crosslink Doublet Distances:* Define any additional doublet distances between the light and one or more heavy crosslinker fragments here. Note that any given doublet must have the light fragment specified in the crosslinker definition set as the light fragment while the heavy fragment can vary. Values set here will be parsed and available for selection in the *Additional Crosslink Doublet Distances* setting for MS Annika Detector.
- *Diagnostic Ions:* See *Diagnostic ions* above. Set potential values for selection here as a semicolon separated list. After reloading open studies, these masses should be available for selection in the Detector node.

MS Annika Search

The MS Annika Search Engine node is very similar to the MS Amanda database search node. Usual search engine parameters can be set here. For more information see the MS Amanda web page. An MS Annika CSM Validator is needed to calculate an FDR and persist CSMs and cross-links.

- *Protein Database:* The FASTA database to use. Provide it in *Administration > Maintain FASTA Files*.
- *Enzyme name:* The enzyme used for *in-silico* digestion of protein databases.
- *MS1 Tolerance:* The tolerance used to compare data at the MS1 level.
- *MS2 Tolerance:* The tolerance used to compare data at the MS2 level.

- *Missed Cleavages:* The number of missed cleavages considered in the digest of the protein database.
- *Search Approach:* Search for crosslinks in MS2 spectra or search for crosslinks in MS2 *and* MS3 spectra.
- *Precursor Selection:* This should match the setting selected in the 'Spectrum Selector Node'. (Only affects MS3 search)
- *MS3 tolerance:* Fragment ion tolerance used at the MS3 level. (Only affects MS3 search)
- *Create MS2 spectra:* If set to 'True' the peptide masses inferred from the MS3 spectra will be used to also search the precursor MS2 spectrum to identify the peptides. Otherwise, the MS2 spectra will only be searched for peptide masses directly found in the MS2 spectra. (Only affects MS3 search)
- *Multiplicative Penalty for Crosslinker with equal Sequences:* Scores of CSMs with two equal sequences will be multiplied by this value. To negate the effect of this parameter, set the value to 1. If set to a value greater than 1, CSMs containing two identical peptides are preferred.
- *Perform Decoy Search:* Whether CSM Decoy Search should be performed.

Advanced parameters for MS3 search:
- *Boost:* If a peptide is identified in the MS2 precursor and one or more MS3 product ion spectra its score will be boosted by the number of percent specified here (see also publication for the exact scoring function). Usually the default of 20 is sufficient and should not be changed outside of benchmarking when the ground truth is known! (Only affects MS3 search)
- *Isotopic Peak Tolerance:* Maximum allowed deviation of m/z for peaks of the same isotopic envelope e.g., if set to 0.01 a doubly charged peak will still be considered part of the isotopic envelope if it is between 0.49 and 0.51 m/z away from the next peak of the isotopic envelope. (Only affects MS3 search)
- *Maximum Intensity Difference:* Maximum relative intensity difference between peaks of the same isotopic envelope given as the fraction of the next, more intense peak e.g., if set to 0.3 a peak must have at least 30% as much intensity as the neighboring more intense peak to be considered to be of the same isotopic envelope. (Only affects MS3 search)
- *Maximum Considered Precursor Charge:* The maximum charge to be considered during search when the charge of the precursor is unknown e.g., if set to 4 the charges 1, 2, 3, 4 will be considered as possible charge states for the precursor. (Only affects MS3 search)

MS Annika Validator

The MS Annika Validator node allows for the calculation CSM as well as cross-link-level FDR. Furthermore, the Validator is responsible for persisting the CSMs and cross-links to Proteome Discoverer. Found CSMs are attributed with a confidence value according to the settings described below. Then all CSMs are reported in Proteome Discoverer with their respective confidence.

- *Medium confidence FDR cutoff:* Percentage for FDR calculation, usually 5%. CSMs are considered as **decoy** as soon as at least one of the crosslinker peptides is a decoy. The FDR is calculated at the CSM level.
- *High confidence FDR cutoff:* Percentage for FDR calculation, usually 1%.
- *Include Decoy CSMs/Cross-links:* Whether decoy CSMs or cross-links are included in the output. This is included here since there is no Decoy CSM or Decoy Crosslink tab in Proteome Discoverer (yet). The respective items are marked as decoy or target in the output.
- *Separate Intra/Inter-link FDR:* Determines if Intra- and Interlink CSMs are separated before FDR calculation. If set to true, an FDR calculation is done for each subset of CSMs and those two subsets are combined to yield the result.
- *Group Crosslinks by:* Choose how crosslinks are grouped in the output. They can either be grouped by the position of the cross-linker in the protein sequence or the peptide sequence.
- *DeltaCn Filter:* Retain only CSMs with a DeltaScore less than the specified value. This is only relevant when persisting multiple CSMs for one scan number. The delta score is calculated on CSMs sorted by score. The formula is $1-score(csm_n)/score(csm_c)$ where csm_n is the CSM with the lower score, and csm_c the CSM with the

higher score. This score is a measure of similarity ranging between zero and one. The more similar two results, the lower the score.

- *Individual peptide score filter:* A score cutoff. Both CSM scores (for the alpha and beta peptide) must clear this threshold, or the CSM is discarded.
- *Top N CSMs:* The number of CSMs displayed in the output for each scan number. This allows for inspection of multiple CSMs identified from a spectrum.

General Remarks

to work, MS Annika needs the right crosslinker masses. The predefined DSSO linker is in the right format: the light part and the heavy part are supplied. In the case of DSSO, a water loss in the linker is also possible. The additional doublet resulting from this must be specified in the *Additional Crosslink Doublet Distances* parameter in the Detector, for DSSO this value is already defined and can be readily selected, any other additional doublets for other crosslinkers need to be specified in *Administration > Configuration > MS Annika* Detector *> Additional Crosslink Doublet Distances* (see MS Annika Detector). Additional crosslinking sites can be defined in the *Maintain Chemical Modifications* tab. For more information on how to define masses, visit the MS Annika home page.