



Organizing MS/MS Proteomic Data for Publication with Scaffold

Brian C. Searle
Proteome Software Inc.
Portland, Oregon

ABRF2009, Memphis TN
February 10th, 2009



Publication Standards

- In 2006 Molecular and Cellular Proteomics published guidelines for reporting peptide and protein identifications
- Other proteomics journals have adopted similar standards



What do the Guidelines Do?

- The guidelines ensure enough information is provided to:
- Understand and critically assess the results
 - Enforce a low level of reliability of the results
 - Provide enough data concerning potentially questionable results to allow some reassessment



What do the Guidelines Do?

- The guidelines ensure enough information is provided to:
- Understand and critically assess the results
 - Enforce a low level of reliability of the results
 - Provide enough data concerning potentially questionable results to allow some reassessment

Publication guidelines have played a critical role in the acceptance of proteomic analysis



Publication Standards are Hard!

- They're hard for the authors who have to comply



Publication Standards are Hard!

- They're hard for the authors who have to comply
- They're hard for the reviewers who are now tasked with the job of policing compliance



Publication Standards are Hard!

- They're hard for the authors who have to comply
- They're hard for the reviewers who are now tasked with the job of policing compliance
- They're hard for the journals because there's a huge amount of supplemental data in a variety of unstructured formats



Publication Standards are Hard!

- They're hard for the authors who have to comply
- They're hard for the reviewers who are now tasked with the job of policing compliance
- They're hard for the journals because there's a huge amount of supplemental data in a variety of unstructured formats

But they don't have to be!



Making the Guidelines Easier for Authors to Follow

- Collate relevant data out of search engine files
 - Most search engines supported: Mascot, SEQUEST, XI Tandem, Phenyx, SpectrumMill, OMSSA, IdentityE
 - Collapse related parameters across all result files
 - Clearly point out metadata not present in files
 - Produce search engine and instrument independent probabilities for comparison



Collate Relevant Data

- Peak picking software, version, altered parameters
- Database Selection
 - Database name and version
 - Species restriction
 - Number of proteins searched
- Database search parameters
 - Search engine name and version
 - Enzyme specificity
 - # missed cleavages
 - Fixed/variable modifications
 - Mass tolerances
- Peptide selection criteria

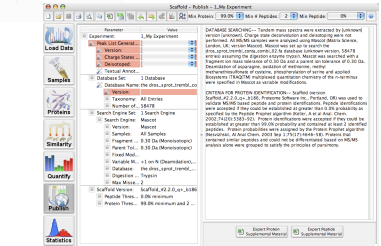


Collate Relevant Data

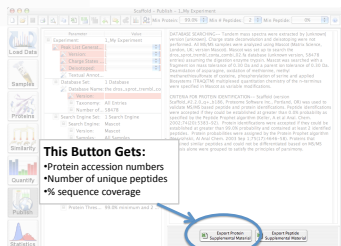
- Peak picking software, version, altered parameters
- Database Selection
 - Database name and version
 - Species restriction
 - Number of proteins searched
- Database search parameters
 - Search engine name and version
 - Enzyme specificity
 - # missed cleavages
 - Fixed/variable modifications
 - Mass tolerances
- Peptide selection criteria



Collate Relevant Data



Collate Relevant Data

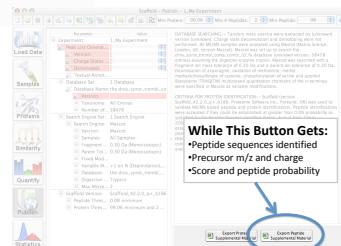


This Button Gets:

- Protein accession numbers
- Number of unique peptides
- % sequence coverage



Collate Relevant Data



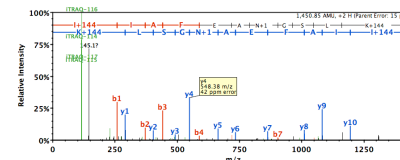
While This Button Gets:

- Peptide sequences identified
- Precursor m/z and charge
- Score and peptide probability



Collate Relevant Data

- Reporting one hit wonders requires releasing the Scaffold file to reviewers to interrogate peak assignments



Making the Guidelines Easier for Reviewers to Police

- Trivial to guarantee that all metadata is present

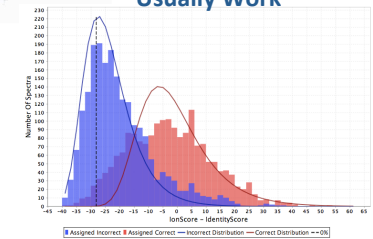
Making the Guidelines Easier for Reviewers to Police

- Trivial to guarantee that all metadata is present
- Quick to review key proteins and spectra

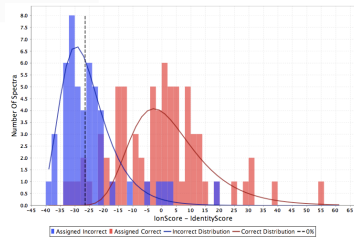
Making the Guidelines Easier for Reviewers to Police

- Trivial to guarantee that all metadata is present
- Quick to review key proteins and spectra
- Easy to validate that statistical assumptions are sound

Probability Assumptions Usually Work



But Not Always!



Distributing Guideline Compliant Data

- Journals can distribute Scaffold files that contain the entire data set

Distributing Guideline Compliant Data

- Journals can distribute Scaffold files that contain the entire data set
- Authors can allow their data to be open
 - Peak list exports for a variety of platform independent formats (MGF, DTA, PKL, etc)

Distributing Guideline Compliant Data

- Journals can distribute Scaffold files that contain the entire data set
- Authors can allow their data to be open
 - Peak list exports for a variety of platform independent formats (MGF, DTA, PKL, etc)
- Journals are allowed to distribute the specific version of the free viewer along with the files so the files can ALWAYS be opened!

Conclusions

Scaffold makes it:

- Easier to pass publishing data guidelines
- Easier to review data and to police standards
- Easier to safely distribute supplemental material in a useful format