Draft Guidelines for Publication of Peptide and Protein Identification Data

Journal of Molecular and Cellular Proteomics Working Group on Publication Guidelines

Steven Carr, Broad Institute of MIT and Harvard (Chair)
Ruedi Aebersold, ETH and Institute for Systems Biology
Michael Baldwin, University of California, San Francisco
Al Burlingame, University of California, San Francisco
Karl Clauser, Broad Institute of MIT and Harvard
Alexey Nesvizhskii, Institute for Systems Biology

Publication Guidelines for Peptide and Protein Identification Data in MCP

Goals:

- try to insure that high quality, significant data are entering the proteomics literature
- develop minimal guidelines for publication of peptide and protein identification data in MCP
- Initial focus on how identifications were made and validated
- guidelines should not be burdensome nor should they dictate what tools to use
- Initiate discussion on requiring submission of data as a condition for acceptance of manuscript and logistics involved

Why are guidelines needed?

Dramatic increase in the number of large data set papers being published

- Lack of accepted and widely available computational tools for reviewers and readers to determine if results are valid
- Published studies often do not contain enough information for the reader to assess how the data was processed and what the criteria for identification were
- Lack of understanding and misuse of algorithms contribute to large false positive error rates
- Likely that we are publishing many incorrect interpretations

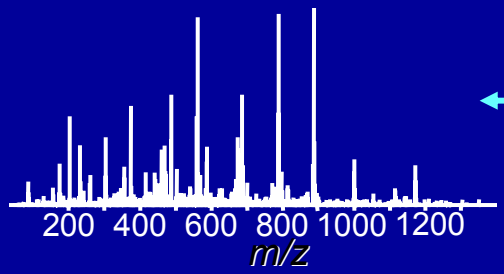
Why are guidelines needed?

- Finding a peptide match in a DB is easy, but knowing whether it is correct is not

MS/MS Database Search

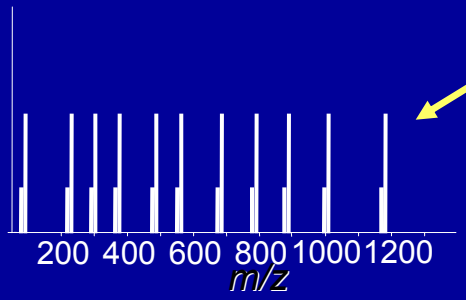
Acquired MS/MS spectrum

Sequence Database



correlate

Theoretical spectrum



ISLLDAQSAPLR
VVEELCPTPEGK
DLLLQWCWENGK
ECDVVSNTIIAEK
GDAVFVIDALNR
VPTPNVSVVDLTNR
SYLFCMENSAEK
PEQSDLRSWTAK

similarity score

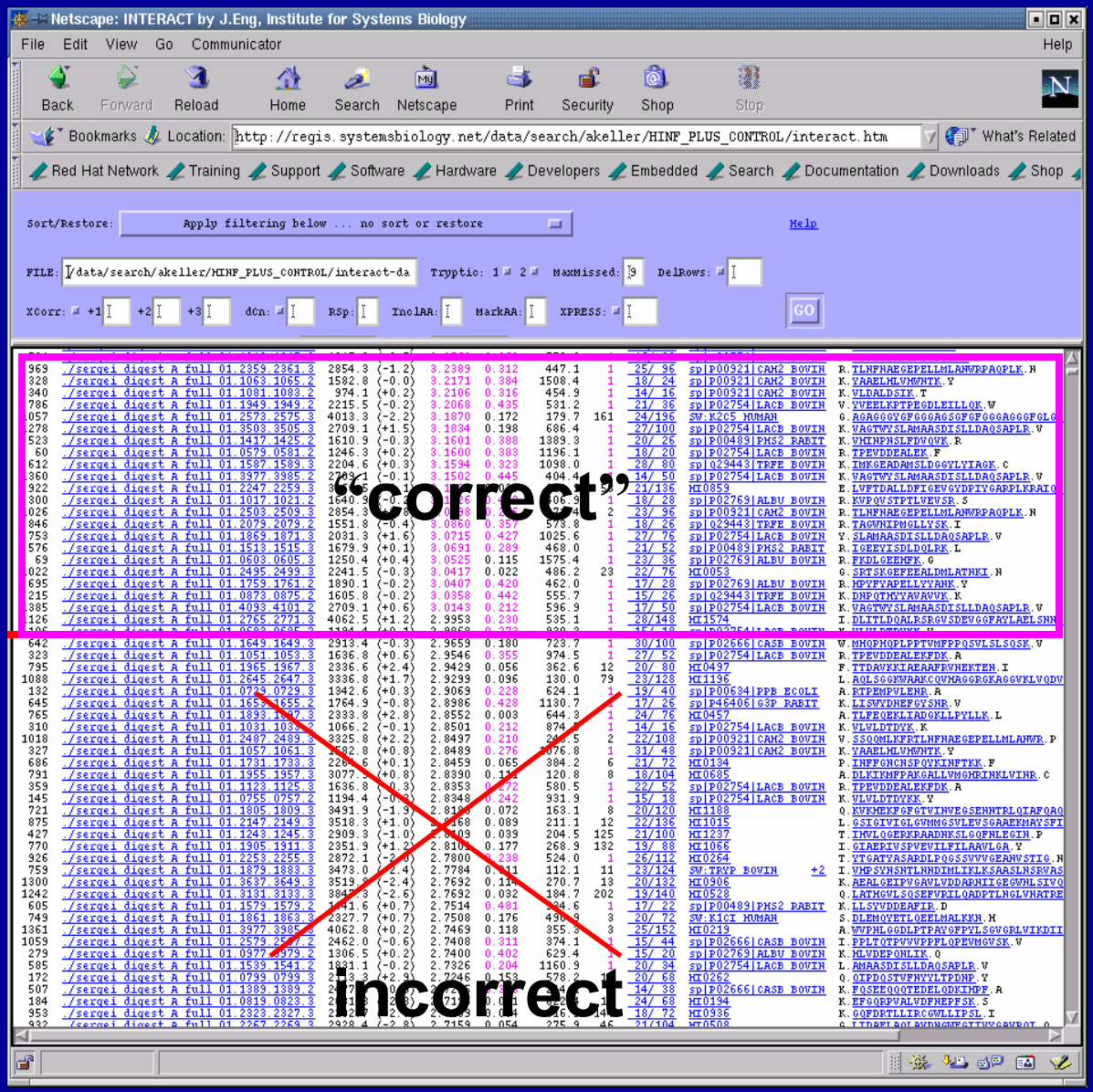
best matching database peptide

Algorithms: SEQUEST, Mascot, Sonar, SpectrumMill, …

best matching peptide in database may be correct or incorrect

Slide courtesy of Alexey Nesvizhskii, ISB

Threshold Model



threshold

SEQUEST:
Xcorr > 2.0
ΔC_n > 0.1

MASCOT:
Score > 30

Slide courtesy of Alexey Nesvizhskii, ISB

Why are guidelines needed?

- Finding a peptide match in a DB is easy, but knowing whether it is correct is not
  - It is almost always possible to match a MS/MS spectrum to a peptide in the database
  - Incorrect matches often (but not always) result from use of low quality peptide MS/MS data to search the database
  - Even high quality data can produce invalid identifications
    - actual peptide sequence is not in the database searched (under the search conditions used)

Why are guidelines needed?

- Unknown and variable false positive error rates are associated with each algorithm
  - Commercial algorithms uses thresholds and scoring methods to move most probable hit to top of list
  - Recommended settings are empirically derived and are not universally applicable
  - Use of conservative scoring and filtering thresholds reduces number of misassigned peptides and proteins, but does not eliminate false positives
  - Probability of a false positive assignment is much higher for “one-hit-wonders”
- statistical methods to validate peptide assignments to MS/MS spectra of peptides have shown promising results, but are not yet widely available or accepted

Publication Guidelines for Peptide and Protein Identification Data in MCP

Working group assembled January, 2004

- Ruedi Aebersold, ETH Zurich and Institute for Systems Biology
- Michael Baldwin, University of California, San Francisco
- Al Burlingame, University of California, San Francisco
- Steven Carr, Broad Institute of MIT and Harvard (Chair)
- Karl Clauser, Broad Institute of MIT and Harvard
- Alexey Nesvizhskii, Institute for Systems Biology
- Additional contributions from: Robert Chalkley, Kirk Hansen, Kati Medzihradszky, UCSF; Andrew Keller, ISB and Ron Beavis, Beavis Informatics, Ltd.

Guidelines published Mol. Cell. Proteomics June 2004; 3: 531.

Guideline 1

Describe search engine used and how peptide and protein assignments were made using that software

All papers must provide:

- The method and/or program used to create the “peak list” from raw data
  - note factors that affect the quality of the subsequent database search (e.g., smoothing, de-isotoping)
- Name and version of DB search program used and parameters used for its operation
  - include precursor-ion mass accuracy; fragment-ion mass accuracy; modifications allowed for; enzyme specified or not; any missed cleavages; etc.

Guideline 1, con't.

- Name and version of sequence database used
  - Include number of protein entries at time of search
- Scores used to interpret MS/MS data
- Thresholds and values specific to judging certainty of identification and description of how applied
- Describe any statistical analysis that was applied to validate the results and of how it was applied
  - e.g. reverse database search

Guideline 2

Provide sequence coverage observed for each protein identified

- the total number of peptides belonging to each protein must be explicitly stated (not # of MS/MS spectra)
- different forms of the same peptide are to be counted as only a single peptide
  - Differing charge states of same peptide or common sample handling artifacts (e.g., ox) all count as 1
- encourage providing tables that list sequences of all identified peptides/protein

Guidelines 3 and 4

Increase the stringency of information required to use single peptide identifications for protein assignment

Protein assignments based on single peptide assignments must include:

- the sequence of the peptide used to make each such assignment, together with the amino acids N- and C-terminal to that peptide's sequence
- the precursor mass and charge (not just m/z) observed
- the scores for this peptide

Guidelines 3 and 4, con't.

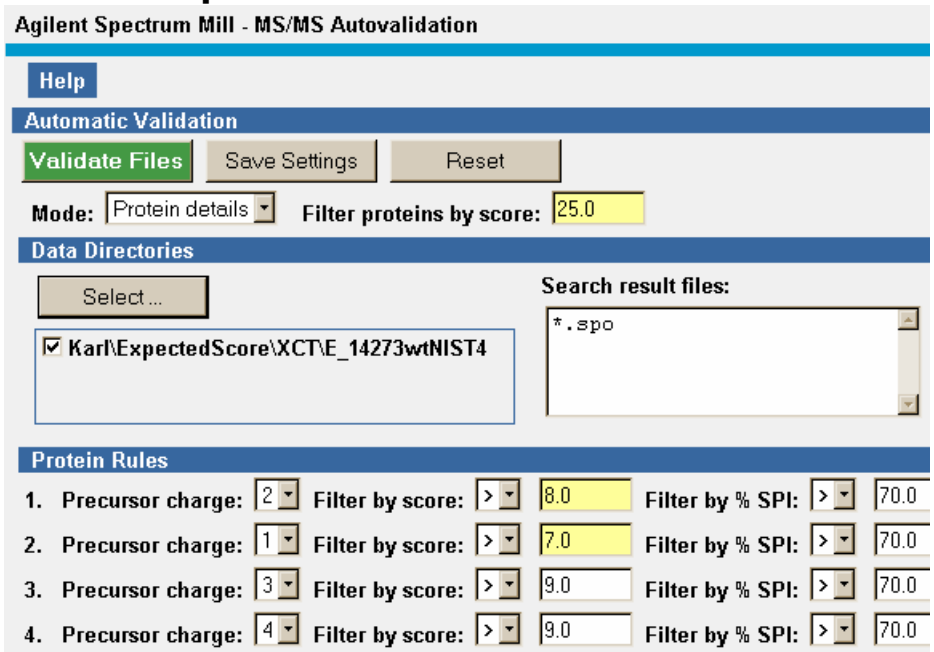
- Biological conclusions based on a single peptide id's or to a posttranslationally modified form of that protein, must be supported by inclusion of the MS/MS spectrum

- Single peptides from ICAT and similar experiments are covered by this guideline as well
  - For large ICAT datasets we have not yet required that spectra for all single-peptide id's be provided

We Use Separate Thresholds for 1-hit Wonders

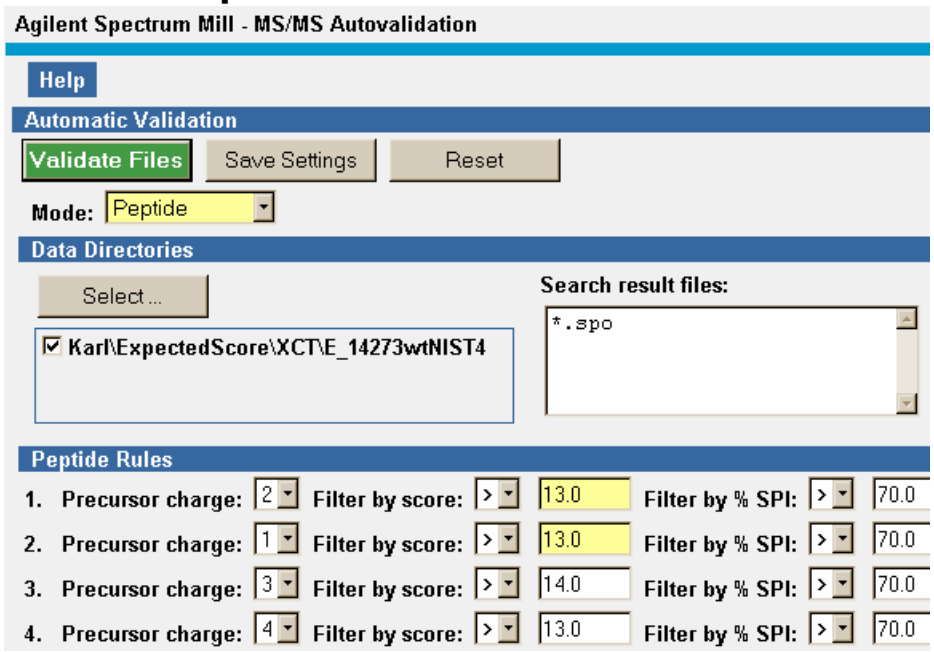
Step 1 - Protein Mode

- 2 or more peptides/protein
- Each spectrum: moderate or better score

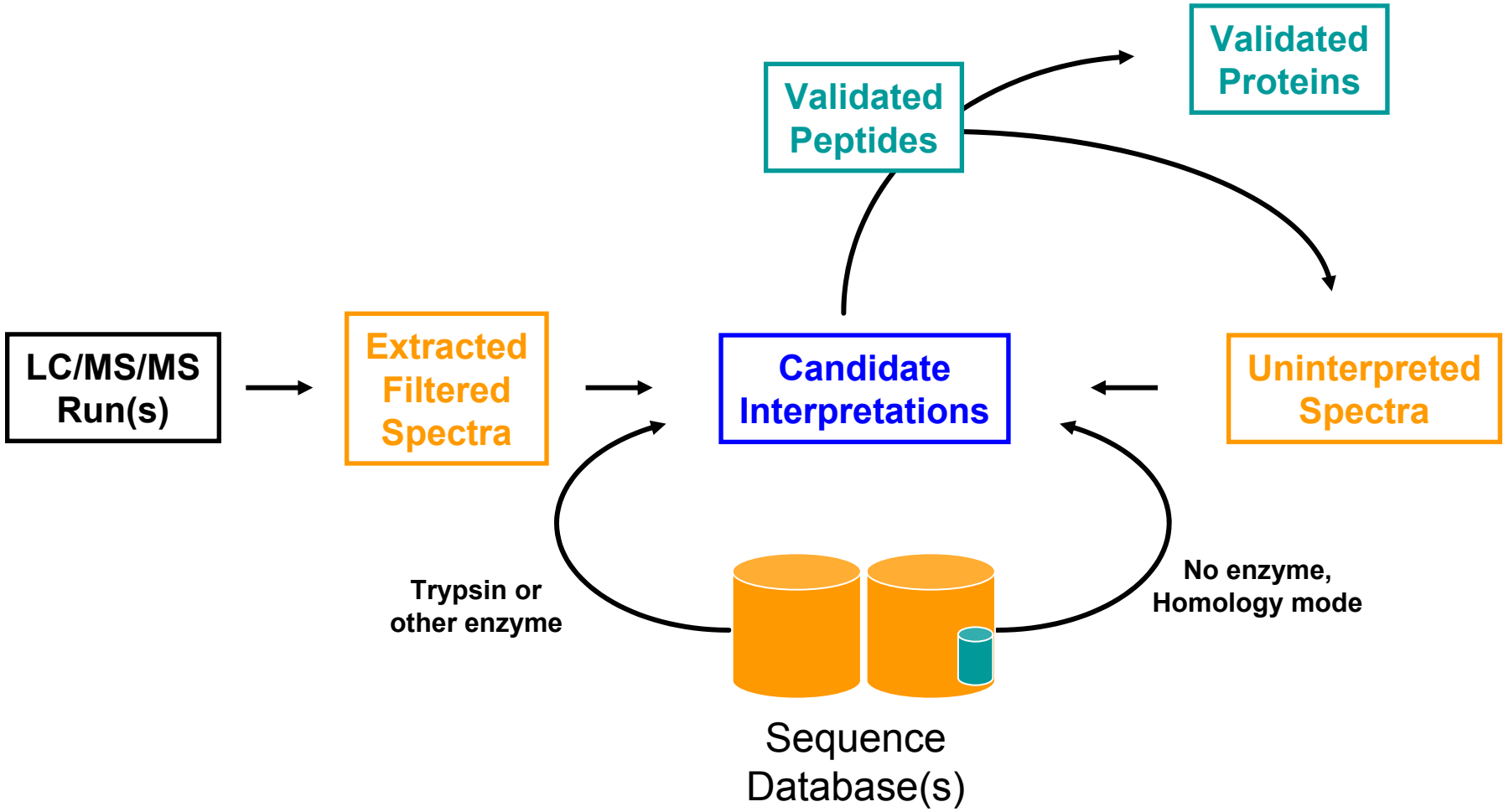


Step 2 - Peptide Mode

- 1 peptide/protein
- Each spectrum: excellent score

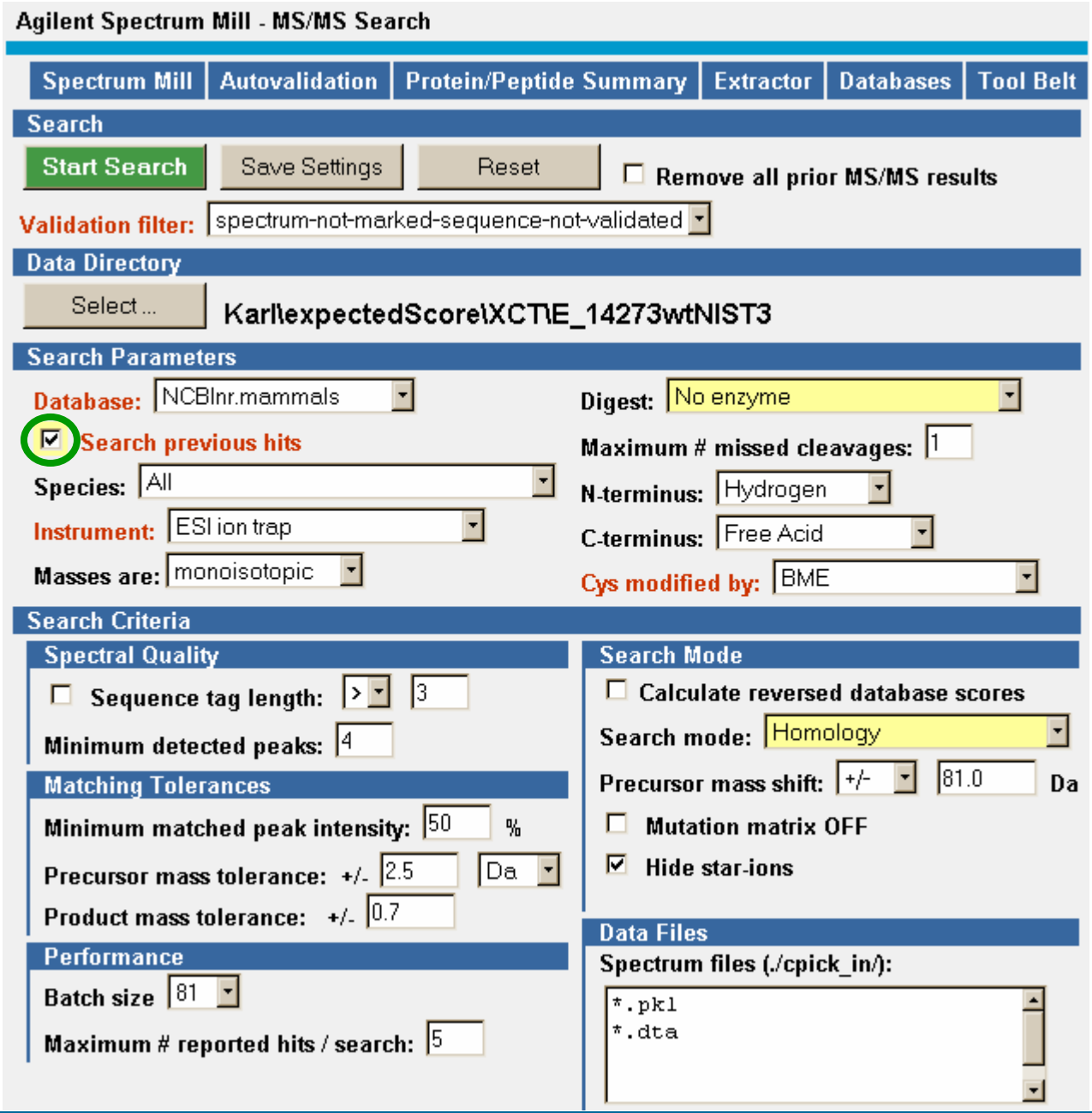


Disallow 1 Hit Wonders that are partial/non-tryptic



No enzyme and homology mode searching of remaining spectra

Using only the subset of proteins confidently identified from a previous trypsin search of the full database

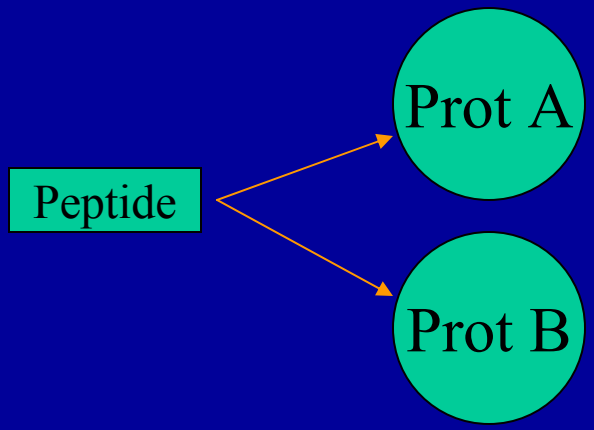


Guideline 6

How to count the number of unique proteins identified based on the peptides found

Protein Inference Problem

Degenerate peptides: correspond to more than a single entry in protein database



protein A or protein B ??

Or both?

In shotgun proteomics the connectivity between peptides and proteins is lost

Degenerate peptides are more prevalent with databases of higher eukaryotes due to the presence of:

- related protein family members
- alternative splice forms
- partial sequences

Slide courtesy of Alexey Nesvizhskii, ISB

Guideline 6

How to count the number of unique proteins identified based on the peptides found

Issue: same (or very similar) protein having different names and accession numbers in the database

– Authors must demonstrate that they are aware of the problem and have taken reasonable measures to eliminate redundancy

– When a single protein member of a multi-protein family has been singled out, explain how the other members of the group were ruled out, if at all

– If a protein from a different species than that studied is identified, then this must be mentioned and justified

Guideline 5

Peptide mass fingerprint data will continue to be accepted for peptide identification, but the standard of acceptability will be more stringent

- list the number of masses matched to the identified protein and the sequence coverage observed
- State the number of masses NOT matched
- Describe parameters and thresholds used to analyze the data (e.g., mass accuracy, res., how calibrated, etc.)
- Authors are encouraged to use and provide the results of scoring schemes which give measure of certainty of id, or perform some measure of false-positive rate

Guideline 7

MCP strongly encourages (but does not at present require) the submission of all MS/MS spectra mentioned in the paper as supplemental material.

- We will accept dta, pkl, mgf files

MCP is moving toward accepting and serving raw or minimally processed MS data, but we are not there yet

- Technical aspects of storing large repositories of raw mass spectrometric data has yet to be worked out
- Authors are encouraged to provide access to raw MS data using group websites etc.
  - **Not a viable, long-term solution. Public repositories are essential.**

Capacity Constraints on Repositories

File type	LCQ-Deca (centroid)	LTQ (centroid)	LTQ-FT (centroid)	QStar (profile)	Qtof (profile)
original/raw (MB)	15	65	200	75	500
Winzip compresses to (%)	71	83	83	50	50

Current/Future Utility Constraints on Readers/Reviewers

- Lowest common denominator currently is the original instrument vendor format.
    - Files contain all the interesting info in unprocessed form
        - parent peak intensities for quantitation
        - acquisition parameters

However…

- If repository stores original instrument vendor format, user needs instrument vendor's data system to read files

Current/Future Utility Constraints on Readers/Reviewers

- If repository stores XML format, then user needs compatible tools
  - –ISB provides converters from most instruments to mzXML and open source non-graphical mzXML reader
  - –mzData - similar XML format from HUPO, but no converters or readers available yet
- Will search engines support XML files?
- Will Instrument vendors formats continue to be compatible with XML converters?
  - Meetings like this need to have representatives from MS manufacturers present who are in decision-making capacity
- Will open source community provide viable graphical utilities for XML formats?
- Will they work on decreasing dataset size?

Next Steps

Meeting devoted to publication guidelines for proteomics data and data repository issues

Goals:

- to come up with an agreed upon set of standards for proteomics data publication/presentation
- to develop clear and actionable plans for data sharing with testable mechanisms to be put into place in 2005
  - journal editors
  - Tool developers
  - Instrument vendors
  - Power users

Coordinate with PSI-HUPO and other serious groups

Acknowledgements

Karl Clauser, Broad Institute

Ralph Bradshaw, UC Irvine (Editor, MCP)

Barbara Gordon, MCP

Highwire Press, Stanford University

Over 15 million articles from over 4,500 PubMed journals, including 819,165 free full text articles from 779 HighWire-hosted journals