

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

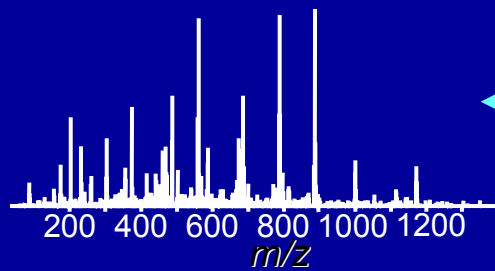
- Lack of accepted and widely available computational tools for reviewers and readers to determine if results are valid
- Published studies often do not contain enough information for the reader to assess how the data was processed and what the criteria for identification were
- Lack of understanding 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

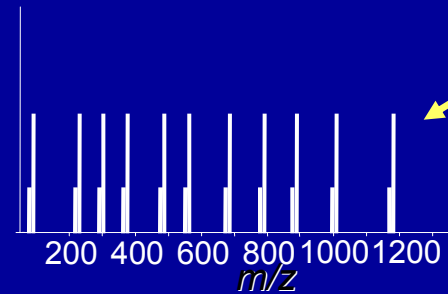
MS/MS Database Search

Acquired MS/MS
spectrum



Sequence Database

Theoretical spectrum



ISLLDAQSAPLR
VVEELCPTPEGK
DLLLQWCWENGK
ECDVVSNTIIAEK
GDAVFVIDALNR
VPTPNVSVVDLTNR
SYLFCMENSAEK
PEQSDLRSWTAK

correlate

similarity score

best matching database peptide

Algorithms: SEQUEST, Mascot, Sonar, SpectrumMill, ...

best matching peptide in database may be correct or incorrect

Threshold Model

sort by search score

Netscape: INTERACT by J.Eng, Institute for Systems Biology

Location: http://regis.systemsbiology.net/data/search/akeller/HINF_PLUS_CONTROL/interact.htm

Sort/Restore: Apply filtering below ... no sort or restore

FILE: /data/search/akeller/HINF_PLUS_CONTROL/interact-da

Xcorr: +1 +2 +3 dcn: RSP: InclRA: MarkRA: XPRESS:

969	/sergei digest A full 01 2359 2361.3	2854.8	(-1.2)	3.2389	0.312	447.1	1	25/36	SP P00921 CAM2 BOVIN	R. TLNFNAEGPELLMLANVPAQPLK.N
328	/sergei digest A full 01 1063 1055.2	1582.8	(-0.0)	3.2171	0.384	1508.4	4	18/24	SP P00921 CAM2 BOVIN	K. VAREMLLHVHNTK.V
340	/sergei digest A full 01 1081 1083.2	974.1	(+0.2)	3.2106	0.316	454.9	1	14/16	SP P00921 CAM2 BOVIN	K. VLRLDLSDK.T
796	/sergei digest A full 01 1949 1949.2	2215.5	(-0.2)	3.2068	0.435	531.2	1	21/36	SP P02754 LACB BOVIN	V. YVEELKPTTPEGDLLELLOK.W
057	/sergei digest A full 01 3575 3575.3	4013.3	(+0.2)	3.1870	0.192	179.7	161	24/100	SW-K265 HUMAN	G. HORAGGGGFGGAGGGGFGGAGGGGFGG
278	/sergei digest A full 01 3503 3505.3	2709.1	(+1.5)	3.1834	0.198	686.4	1	27/100	SP P02754 LACB BOVIN	K. VAGTQYSLAWASDLSLDAGSAPLV.V
523	/sergei digest A full 01 1417 1425.2	1610.9	(-0.3)	3.1601	0.388	1389.3	1	20/26	SP P00489 PHS2 RABIT	K. VGHINPHSLPDUQK.R
60	/sergei digest A full 01 0579 0581.2	1246.3	(+0.2)	3.1600	0.389	1196.1	1	18/20	SP P02754 LACB BOVIN	R. TPFEVDEALEK.F
612	/sergei digest A full 01 1587 1589.3	2204.6	(+0.3)	3.1594	0.323	1098.0	1	28/30	SP Q29443 TRFE BOVIN	K. IUKGEGADAMSLDGGVLYIACK.C
360	/sergei digest A full 01 3977 3985.2	2799.1	(-0.1)	3.1502	0.445	404.4	1	14/50	SP P02754 LACB BOVIN	K. VAGTQYSLAWASDLSLDAGSAPLV.V
922	/sergei digest A full 01 2293 2293.3	2799.1	(-0.1)	3.1502	0.445	404.4	1	23/106	MI0359	E. IYTRDLRFLPSEVCHDPTTFCARGLPKALIO
300	/sergei digest A full 01 1017 1021.2	1640.9	(-0.3)	3.1498	0.388	1063.9	1	18/28	SP P02769 ALBU BOVIN	R. KVPDUSTPTLVEUSE.S
926	/sergei digest A full 01 2503 2503.3	2854.8	(-0.0)	3.0886	0.357	573.8	2	23/36	SP P00921 CAM2 BOVIN	R. TLNFNAEGPELLMLANVPAQPLK.N
846	/sergei digest A full 01 2079 2079.2	1551.8	(-0.4)	3.0860	0.357	573.8	1	18/26	SP Q29443 TRFE BOVIN	R. TAGOHIFMGLYSK.I
753	/sergei digest A full 01 1859 1871.3	2091.3	(+1.6)	3.0715	0.427	1025.6	1	27/76	SP P02754 LACB BOVIN	V. SLAWASDLSLDAGSAPLV.V
576	/sergei digest A full 01 1519 1519.3	1679.9	(+0.1)	3.0691	0.269	468.0	1	24/52	SP P00489 PHS2 RABIT	R. KQEVITSDGLK.L
389	/sergei digest A full 01 0603 0605.3	1250.4	(+0.4)	3.0525	0.115	1575.4	1	31/48	SP P02769 ALBU BOVIN	R. FKLQGEHMK.G
022	/sergei digest A full 01 2495 2499.3	2241.5	(-0.3)	3.0417	0.022	486.2	23	22/76	MI0053	G. SETSKGFFFEALDMLATNKI.N
635	/sergei digest A full 01 1759 1761.2	1890.1	(-0.2)	3.0407	0.420	462.0	1	17/28	SP P02769 ALBU BOVIN	R. HPTFYAPPELLYVANK.V
215	/sergei digest A full 01 0873 0875.2	1605.8	(-0.2)	3.0358	0.442	555.7	1	15/26	SP Q29443 TRFE BOVIN	K. DNPOTHVAVAVUK.K
395	/sergei digest A full 01 4038 4101.2	2709.1	(+0.6)	3.0148	0.212	596.9	1	17/50	SP P02754 LACB BOVIN	K. VAGTQYSLAWASDLSLDAGSAPLV.V
126	/sergei digest A full 01 2295 2291.3	4052.5	(+1.2)	2.9959	0.230	595.1	1	28/48	MI1524	D. IITLIDDLRLEKQGVSEVGSFVLAELSNH
406	/sergei digest A full 01 0693 0699.3	1194.5	(-0.1)	2.9858	0.293	890.3	1	15/36	SP P02754 LACB BOVIN	V. HAAASDHWASD.V
642	/sergei digest A full 01 1649 1649.3	2913.4	(-0.3)	2.9659	0.180	723.7	1	30/100	SP P02666 CABE BOVIN	W. WHOPHOPFPPTMFPFQVLSLSQSK.W
323	/sergei digest A full 01 1051 1053.3	1636.8	(+0.6)	2.9546	0.355	374.5	1	27/52	SP P02754 LACB BOVIN	R. TPFEVDEALEKFDK.A
795	/sergei digest A full 01 1955 1967.3	2386.6	(+2.4)	2.9429	0.056	362.6	12	20/30	MI0437	F. TTRVYVYIARAFVONKENTN.I
1098	/sergei digest A full 01 2645 2647.3	3936.8	(+1.7)	2.9299	0.096	190.0	79	23/103	MI1156	L. AOLSGGVAKGQUMHAGGKAGGKGLVODV
132	/sergei digest A full 01 0728 0729.3	1342.8	(+0.3)	2.9089	0.228	624.1	1	13/40	SP P00534 PBE ECOLI	R. RYEMPHLPHK.A
645	/sergei digest A full 01 1651 1655.2	1764.9	(-0.8)	2.8986	0.428	1130.7	1	17/26	SP P464061 GSP RABIT	K. LISVYDHEFGYSNR.V
765	/sergei digest A full 01 1838 1840.3	2338.8	(+2.8)	2.8552	0.003	644.3	1	24/76	MI0457	A. TLPEQKLLADCKLLPLLL.L
310	/sergei digest A full 01 1031 1033.3	1066.2	(-0.1)	2.8501	0.212	874.5	1	14/16	SP P02754 LACB BOVIN	K. VLVLDTYK.K
1018	/sergei digest A full 01 2437 2439.3	3325.8	(+2.2)	2.8497	0.210	566.5	2	22/108	SP P00921 CAM2 BOVIN	V. SSSOHLFFPFLNPNFNAEGPELLMLANV.P
327	/sergei digest A full 01 1059 1061.3	3692.8	(+0.8)	2.8489	0.276	376.9	1	31/48	SP P00921 CAM2 BOVIN	K. VAREMLLHVHNTK.V
686	/sergei digest A full 01 1731 1733.3	2287.6	(+0.1)	2.8459	0.065	384.2	6	21/72	MI0134	P. TDFFGNCSPOYKINFTKK.F
791	/sergei digest A full 01 1955 1957.3	3077.6	(+0.8)	2.8390	0.151	120.8	8	18/104	MI0685	A. DLKIMFPKAGALLVMGHEKMLVNR.C
145	/sergei digest A full 01 1123 1125.3	1636.8	(-0.3)	2.8353	0.92	590.5	1	22/52	SP P02754 LACB BOVIN	R. TPFEVDEALEKFDK.A
359	/sergei digest A full 01 0755 0757.2	1194.4	(-0.8)	2.8348	0.242	931.9	1	15/18	MI1368	K. VLVLDTYK.V
721	/sergei digest A full 01 1905 1909.3	3491.9	(-1.2)	2.8169	0.072	163.8	8	20/100	MI1368	Q. VUHCYKFSFGTUNEGSEHNPLQIFPQAO
875	/sergei digest A full 01 2147 2149.3	3518.3	(+1.0)	2.8168	0.089	211.1	12	22/136	MI1016	L. SSGIVTGLGMMHGSVLEVSAGAEKMYSPH
427	/sergei digest A full 01 1243 1245.3	2909.3	(-1.0)	2.8109	0.099	204.5	125	21/100	MI1237	T. THVLQGEKFRADNKLQGNLEGIN.P
770	/sergei digest A full 01 1905 1911.3	2351.9	(+1.2)	2.8101	0.177	268.9	132	19/38	MI1066	I. IAGRIYSPVUVILFLAULGA.V
926	/sergei digest A full 01 2253 2255.3	2872.1	(-0.0)	2.7800	0.338	524.0	1	26/112	MI0254	T. TCGATVASARDLPQGSVVVGEANSTTG.N
759	/sergei digest A full 01 1839 1839.3	3478.0	(-2.4)	2.7784	0.164	112.1	11	23/124	SW-FRPP BOVIN	+2
1300	/sergei digest A full 01 3639 3649.3	3518.3	(-0.6)	2.7408	0.311	374.1	13	19/42	MI1306	T. VDFSYNLSLNDMLTKLSASLNSKVSAS
1242	/sergei digest A full 01 3131 3133.3	3847.8	(-2.6)	2.7632	0.032	184.7	202	19/140	MI0528	K. RELQFQYVGLVADPQNTIQDMLSTVQ
605	/sergei digest A full 01 1579 1579.2	1411.6	(+0.7)	2.7514	0.481	34.6	1	17/22	SP P00489 PHS2 RABIT	K. LLSYVDEAFIR.D
749	/sergei digest A full 01 1861 1863.3	2327.7	(+0.7)	2.7508	0.176	49.9	3	20/72	SW-K1CT HUMAN	S. DLEMGVTEIQLMALKKN.H
1261	/sergei digest A full 01 3977 3985.2	4062.8	(+0.2)	2.7469	0.118	355.3	3	25/152	MI0219	R. VVFNLGGDLPTPYGFFPLSGVGLVWIKDII
1059	/sergei digest A full 01 2579 2624.2	2452.0	(-0.6)	2.7408	0.311	374.1	13	19/42	SP P02666 CABE BOVIN	I. FLTQYTHVPELLOPEVMSVSK.S
279	/sergei digest A full 01 0927 0929.2	1306.5	(+0.2)	2.7400	0.402	623.4	4	15/20	SP P02769 ALBU BOVIN	K. HLYPDEONLTK.Q
585	/sergei digest A full 01 1539 1541.2	1831.1	(-0.2)	2.7326	0.204	1160.9	1	20/34	SP P02754 LACB BOVIN	L. AWASDLSLDAGSAPLV.V
172	/sergei digest A full 01 0799 0799.3	2498.9	(+2.9)	2.7246	0.153	578.8	1	20/68	MI0262	Q. QIPDQSTVFVNVILTPDNE.V
507	/sergei digest A full 01 1839 1839.2	2498.9	(+2.9)	2.7246	0.153	578.8	1	14/38	SP P02666 CABE BOVIN	R. FQSEPOQTEDELQDKTIFP.A
104	/sergei digest A full 01 0819 0823.3	2498.9	(+2.9)	2.7246	0.153	578.8	1	24/68	MI0134	K. FQSEPOQTEDELQDKTIFP.A
953	/sergei digest A full 01 2293 2293.3	2498.9	(+2.9)	2.7246	0.153	578.8	1	24/68	MI0134	K. FQSEPOQTEDELQDKTIFP.A
932	/sergei digest A full 01 2293 2293.3	2498.9	(+2.9)	2.7246	0.153	578.8	1	24/68	MI0134	K. FQSEPOQTEDELQDKTIFP.A

correct

incorrect

SEQUEST:
 $Xcorr > 2.0$

$\Delta C_n > 0.1$

MASCOT:
Score > 30

← threshold

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

Agilent Spectrum Mill - MS/MS Autovalidation

Help

Automatic Validation

Validate Files Save Settings Reset

Mode: Protein details Filter proteins by score: 25.0

Data Directories

Select ... Search result files:

Karl\ExpectedScore\XCTE_14273wtNIST4 *.spo

Protein Rules

1. Precursor charge:	2	Filter by score:	>	8.0	Filter by % SPI:	>	70.0
2. Precursor charge:	1	Filter by score:	>	7.0	Filter by % SPI:	>	70.0
3. Precursor charge:	3	Filter by score:	>	9.0	Filter by % SPI:	>	70.0
4. Precursor charge:	4	Filter by score:	>	9.0	Filter by % SPI:	>	70.0

Step 2 - Peptide Mode

- 1 peptide/protein
- Each spectrum: excellent score

Agilent Spectrum Mill - MS/MS Autovalidation

Help

Automatic Validation

Validate Files Save Settings Reset

Mode: Peptide

Data Directories

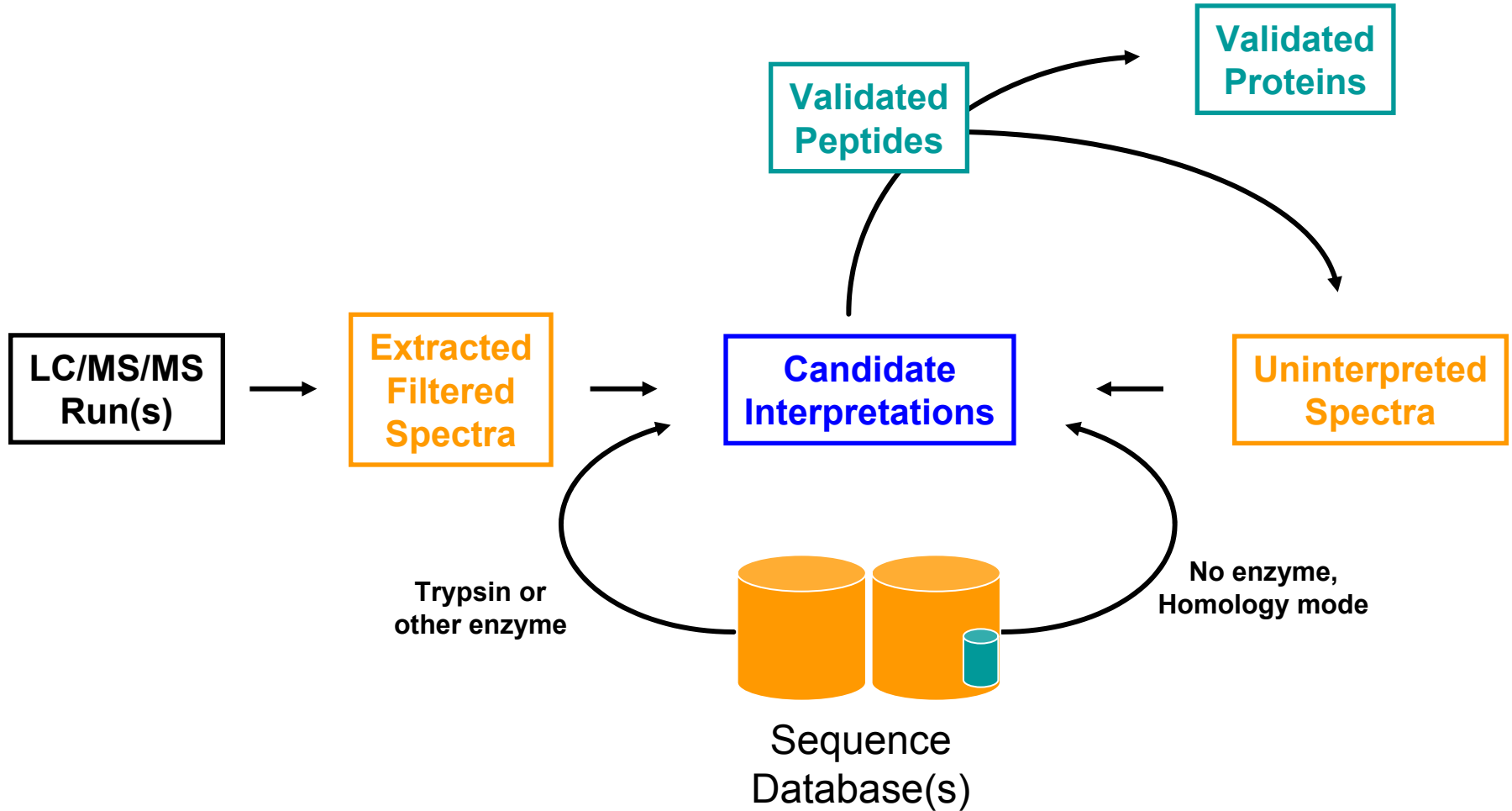
Select ... Search result files:

Karl\ExpectedScore\XCTE_14273wtNIST4 *.spo

Peptide Rules

1. Precursor charge:	2	Filter by score:	>	13.0	Filter by % SPI:	>	70.0
2. Precursor charge:	1	Filter by score:	>	13.0	Filter by % SPI:	>	70.0
3. Precursor charge:	3	Filter by score:	>	14.0	Filter by % SPI:	>	70.0
4. Precursor charge:	4	Filter by score:	>	13.0	Filter by % SPI:	>	70.0

Disallow 1 Hit Wonders that are partial/non-tryptic



Agilent Spectrum Mill - MS/MS Search

Spectrum Mill | Autovalidation | Protein/Peptide Summary | Extractor | Databases | Tool Belt

Search

Start Search

Save Settings

Reset

Remove all prior MS/MS results

Validation filter: spectrum-not-marked-sequence-not-validated

Data Directory

Select...

Karl\expectedScore\XCTIE_14273wtNIST3

Search Parameters

Database: NCBI\nr.mammals

Digest: No enzyme

Search previous hits

Maximum # missed cleavages: 1

Species: All

N-terminus: Hydrogen

Instrument: ESI ion trap

C-terminus: Free Acid

Masses are: monoisotopic

Cys modified by: BME

Search Criteria

Spectral Quality

Sequence tag length: > 3

Minimum detected peaks: 4

Matching Tolerances

Minimum matched peak intensity: 50 %

Precursor mass tolerance: +/- 2.5 Da

Product mass tolerance: +/- 0.7

Performance

Batch size: 81

Maximum # reported hits / search: 5

Search Mode

Calculate reversed database scores

Search mode: Homology

Precursor mass shift: +/- 81.0 Da

Mutation matrix OFF

Hide star-ions

Data Files

Spectrum files (./cpick_in/):

*.pk1
*.dta

No enzyme and homology mode searching of remaining spectra

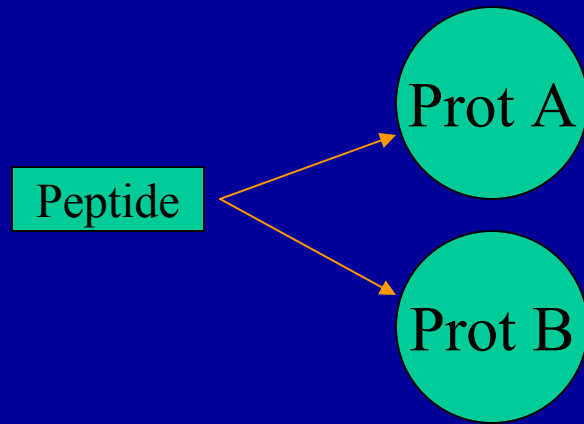
Using only the subset of proteins confidently identified from a previous trypsin search of the full database

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

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