Quantitation and Validation
"Can We Believe These Results?"

Gordon Anderson

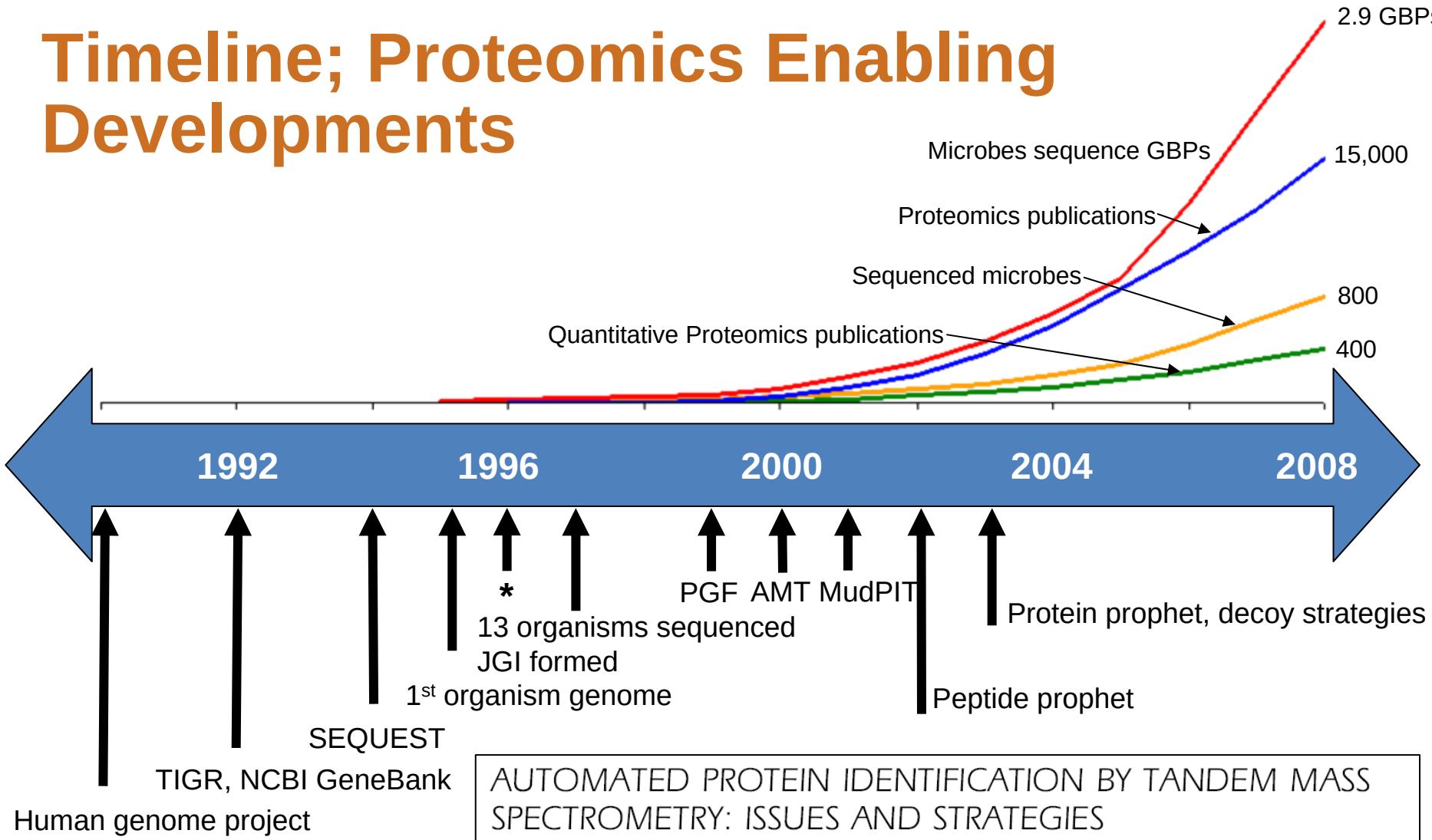
Biological Sciences Division
Fundamental and Computational Science Directorate
Pacific Northwest National Laboratory

February 9, 2009

Objectives

- Overview of Mass Spectrometry Techniques in Quantitative Proteomics and Metabolomics
- Timeline of Enabling Developments
- Proteomics and Metabolomics Challenges
  - Methodologies
  - Validation
- “Can We Believe These Results?”

Timeline; Proteomics Enabling Developments



- Separations with Accurate mass MS, 1996

AUTOMATED PROTEIN IDENTIFICATION BY TANDEM MASS SPECTROMETRY: ISSUES AND STRATEGIES

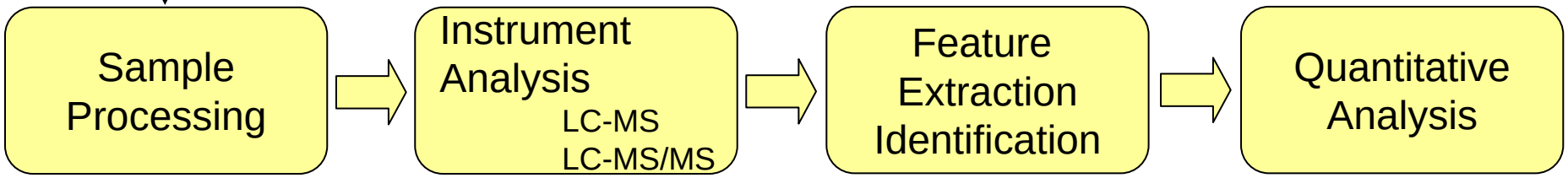
Patricia Hernandez,[1]* Markus Müller,[3] and Ron D. Appel[1,2]
[1]*Swiss Institute of Bioinformatics, Geneva, Switzerland*
[2]*University of Geneva and Geneva University Hospital, Geneva, Switzerland*
[3]*Institute of Molecular Systems Biology, Swiss Federal Institute of Technology, Zürich, Switzerland*

Received 30 November 2004; received (revised) 29 March 2005; accepted 6 April 2005

Published online 11 November 2005 in Wiley InterScience (www.interscience.wiley.com) DOI 10.1002/mas.20068

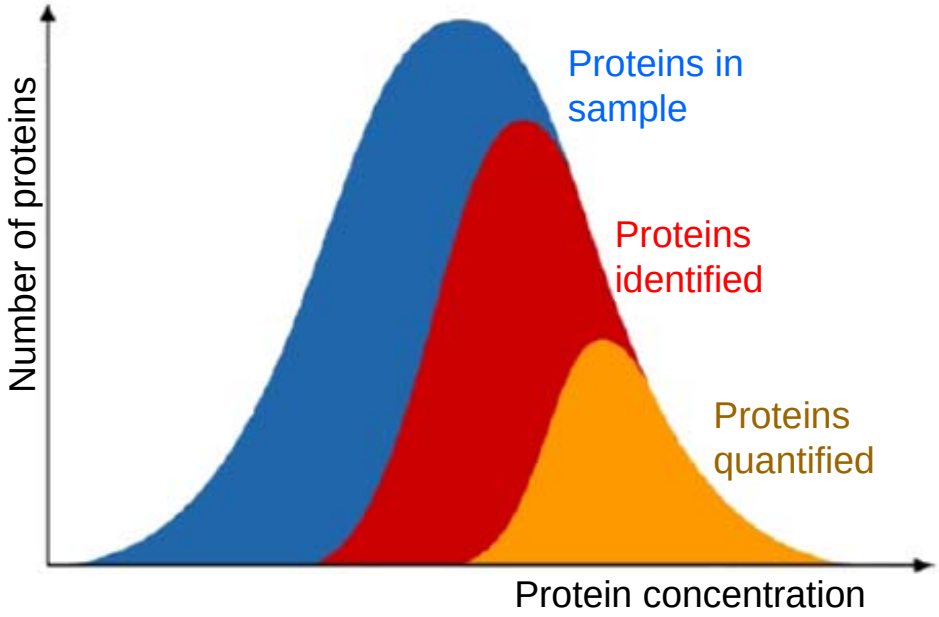
Proteomics Workflow

Cells / Tissue



Purification
Fractionation
Protein extraction
Digestion
Labeling
Spiking

TOF
Ion Traps
Q-TOF
TOF-TOF
FTICR
Orbitrap



Bantscheff, M., M. Schirle, G. Sweetman, J. Rick, and B. Kuster, *Quantitative mass spectrometry in proteomics: a critical review.* Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.

Identification Strategies

- Proteomics
  - MS (Peptide Mass Fingerprinting or PMF)
    - Low complexity mixtures
  - LC-MS/MS (Peptide Fragment Fingerprinting or PFF)
    - Comprehensive tool set available
  - Accurate Mass and Time (AMT) tag approach
    - Requires database of peptide IDs and LC elution times from above
    - High throughput
  - Validation
    - Peptide ID confidence
    - Peptide to protein assignment
    - Protein identification confidence

- Metabolomics
  - Identification tools less mature
    - Accurate mass can be used to determine elemental compositions
    - Structural determination
      - Manual analysis of MS/MS spectra
      - NMR analysis
      - Use of IMS structural information with MS

Pros and Cons of PMF/PFF Strategies

Computational method	Data	Pro/cons
Database dependent PMF search	MS	Accurate Database dependent Fast interpretation Few modifications can be included Only for simple protein mixtures
Database dependent PFF search	MS/MS	Accurate Complex protein mixtures Database dependent Moderate amount of modifications can be included Moderate search speed
Sequence tag + database search	MS/MS	Moderate accuracy[a)] Complex protein mixtures Database dependent Many modifications can be included Fast interpretation
De novo sequencing	MS/MS	Less accurate[a)] Complex protein mixtures Database independent Few modifications can be included Fast interpretation
Spectral library search	MS/MS	Highly accurate Complex protein mixtures Database dependent Few modifications Fast interpretation

a) Likely to become more accurate with future instrument and algorithmic improvements.

Matthiesen, R., *Methods, algorithms and tools in computational proteomics: a practical point of view.* Proteomics, 2007. **7**(16): p. 2815-32.

Quantitation Strategies

- Proteomics
  - Label based (Relative/Absolute)
    - Metabolic labeling
    - Chemical labeling
    - Enzymatic labeling
  - Label free (Relative/Absolute)
  - Peptide to protein “rollup”
    - Degenerate peptide problem
    - Normalization methods
- Metabolomics
  - Primarily label free approaches
  - Does not suffer from the “roll-up” challenge

Quantitation Strategies

Table 1 Characteristics and applications of quantitative mass spectrometry methods

	Application	Accuracy (process)	Quantitative proteome coverage	Linear dynamic range[a]
Metabolic protein labeling	Complex biochemical workflows Comparison of 2–3 states Cell culture systems only	+++	++	1–2 logs
Chemical protein labeling (MS)	Medium to complex biochemical workflows Comparison of 2–3 states	+++	++	1–2 logs
Chemical peptide labeling (MS)	Medium complexity biochemical workflows Comparison of 2–3 states	++	++	2 logs
Chemical peptide labeling (MS/MS)	Medium complexity biochemical workflows Comparison of 2–8 states	++	++	2 logs
Enzymatic labeling (MS)	Medium complexity biochemical workflows Comparison of 2 states	++	++	1–2 logs
Spiked peptides	Medium complexity biochemical workflows Targeted analysis of few proteins	++	+	2 logs
Label free (ion intensity)	Simple biochemical workflows Whole proteome analysis Comparison of multiple states	+	+++	2–3 logs
Label free (spectrum counting)	Simple biochemical workflows Whole proteome analysis Comparison of multiple states	+	+++	2–3 logs

[a] In MRM mode, dynamic range may be extended to 4–5 logs [65]

Bantscheff, M., M. Schirle, G. Sweetman, J. Rick, and B. Kuster, *Quantitative mass spectrometry in proteomics: a critical review.* Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.

Quantitation Strategies

Target	Name of method or reagent	Isotopes
Metabolic stable-isotope labeling		
None	^{15}N-labeling (^{15}N-ammonium salt)	^{15}N
	Stable isotope labeling by amino acids in cell culture (SILAC)	D, ^{13}C, ^{15}N
	Culture-derived isotope tags (CDIT)	D, ^{13}C, ^{15}N
	Bioorthogonal noncanonical amino acid tagging (BONCAT)	No isotope
Isotope tagging by chemical reaction		
Sulfhydryl	Isotope-coded affinity tagging (ICAT)	D
	Cleavable ICAT	^{13}C
	Catch-and-release (CAR)	^{13}C
	Acrylamide	D
	Isotope-coded reduction off of a chromatographic support (ICROC)	D
	2-vinyl-pyridine	D
	N-t-butyliodoacetamide	D
	Iodoacetanilide	D
	HysTag	D
	Solid-phase ICAT	D
	Visible isotope-coded affinity tags (VICAT)	^{13}C, ^{14}C and ^{15}N
	Acid-labile isotope-coded extractants (ALICE)	D
	Solid phase mass tagging	^{13}C
Amines	Tandem mass tag (TMT)	D
	Succinic anhydride	D
	N-acetoxysuccinamide	D
	N-acetoxysuccinamide: In-gel Stable-Isotope Labeling (ISIL)	D
	Acetic anahydride	D
	Proprionic anhydride	D
	Nicotinoyloxy succinimide (Nic-NHS)	D
	Isotope-coded protein labeling (ICPL,Nic-NHS)	D
	Phenyl isocyanate	D or ^{13}C
	Isotope-coded n-terminal sulfonation (ICens) 4-sulphophenyl isothiocyanate (SPITC)	^{13}C
	Sulfo-NHS-SS-biotin and 13C,D3-methyl iodide	^{13}C and D
	Formaldehyde	D
	Isobaric tag for realtive and absolute quantification (iTRAQ)	^{13}C, ^{15}N and ^{18}O
	Benzoic acid labeling (BA part of ANIBAL)	^{13}C
Lysines	Guanidination (O-methyl-isourea) mass-coded abundance tagging (MCAT)	No isotope
	Guanidination (O-methyl-isourea)	^{13}C and ^{15}N
	Quantitation using enhanced sequence tags (QUEST)	No isotope
	2-Methoxy-4,5-1H-imidazole	D
N-terminus protein	Differentially isotope-coded N-terminal protein sulphonation (SPITC)	^{13}C
N-terminus peptide	N-terminal stable-isotope labelling of tryptic peptides (pentafluorophenyl-4-anilino-4-oxobutanoate)	D or ^{13}C
Carboxyl	Methyl esterification	D
	Ethyl esterification	D
	C-terminal isotope-coded tagging using sulfanilic acid (SA)	^{13}C
	Aniline labeling (ANI part of ANIBAL)	^{13}C
Indole	2-nitrobenzenesulfenyl chloride (NBSCI)	^{13}C

Target	Name of method or reagent	Isotopes
Stable-isotope incorporation via enzyme reaction		
C-terminus peptide	Proteolytic ^{18}O-labeling ($H_2{}^{18}O$)	^{18}O
	Quantitative cysteinyl-peptide enrichment technology (QCET)	^{18}O
Absolute quantification		
None	Absolute quantification (AQUA)	D, ^{13}C, ^{15}N
	Multiplexed absolute quantification (QCAT)	D, ^{13}C, ^{15}N
	Multiplexed absolute quantification using concatenated signature (QconCAT)	D, ^{13}C, ^{15}N
	Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA)	D, ^{13}C, ^{15}N
Label-free quantification		
None	XIC-based quantification	No isotope
	Spectrum sampling (SpS)	No isotope
	Protein abundance index (PAI)	No isotope
	Exponentially modified protein abundance index (emPAI)	No isotope
	Probabilistic peptide scores (PMSS)	No isotope

Panchaud, A., M. Affolter, P. Moreillon, and M. Kussmann, *Experimental and computational approaches to quantitative proteomics: status quo and outlook.* J Proteomics, 2008. **71**(1): p. 19-33.

Validation

- Measurement validation
  - Peptide Protein Identification
    - Confidence algorithms
    - Statistical models
  - Quantitation
    - Less mature than identification confidence
- Functional Validation
  - Western blots
  - Gene knockout
  - Protein assays
  - Protein chemistry
- But all measure something different!

Active Software Development to Address Challenges

- Large array of available tools
  - No universal analysis workflow
- Tools support
  - Peptide ID
  - Identification confidence
    - SMART, epic (PNNL active research)
  - Quantitation
    - Polpitiya et al., DAnTE: a statistical tool for quantitative analysis of -omics data. Bioinformatics, 2008. 24(13): p. 1556-1558.
  - Data management / meta data capture
  - Workflow automation

Table 1. Useful programs for data analysis of MS-based proteomics data

MS identification	Validation	Quantitation	Storage
MASCOT [21, 35, 36][a]	MAE [37][a]	VEMS [38, 39][b]	CPAS [40, 41][b]
ProFound [42][b]	AMASS [43][a]	MSinspect [44][b]	PRIDE [45–47][b]
VEMS [38, 39, 48][b]	SALSA [49][a]	PEPPeR [50][b]	dbVEMS [38, 39][b]
Aldente [51][b]	Qscore [52][b]	OpenMS [53][b]	Proteios [54, 55][b]
MS-Fit[b]	PeptideProphet [30][b]	Mzmine [56, 57][b]	GPMDB [58][b]
PeptIdent[b]	Scope [59][c]	SpecArray [60][b]	**Raw data formats**
MS/MS (Direct Database)	EPIR [61][c]	Peplist [60][b]	MzXML [62][b]
VEMS [38, 39, 48][b]	Spider [63][b]	PEAKSQ (BSI)[c]	mzData [64][b]
MASCOT [21, 35, 36][a]	SILVER [65][b]	MSquant (http://msquant.sourceforge.net/)[b]	**Result formats**
Phenyx [31, 34](GENEBIO)[a]		MSight [66][b]	AnalysisXML [46][b]
SEQUEST (Thermo Finnigan)[c]		RelEx [22][b]	ProtXML [67][b]
X!tandem [68][b]		ASAPratio [69][b]	pepXML [67][b]
ProbId [70][b]		**2D-gels**	**Pipelines**
PopITAM [71][b]		Flicker [72][b]	TPP (tools.proteomecenter.org/TPP.php)[b]
OMSSA [73][b]		Melanie (www.gehealthcare.com) [74–76][a]	ProteinScape™ (www.proteinscape.com)[c]
P-mod [77][a]		PDQuest (www.bio-rad.com)[c]	Scaffold (www.proteomesoftware.com)[c]
PLGS (Waters)[c]		DeCyder (www.gehealthcare.com)[c]	TOPP [78][b]
Paragon (ABI)[c,d]		Delta2D (www.decodon.com)[c]	VEMS (http://personal.cicbiogune.es/rmatthiesen/)[c]
Spectral library search		Progenesis (www.nonlinear.com)[c]	PLGS (www.waters.com)[c]
X ! Hunter [58][b]		Proteomweaver (www.definiens.com/www.bio-rad.com)[c]	
Proteotypic Peptide search			
X! P3 [79][b]			
MS/MS (Tag database)			
GutenTag [80][a]			
InsPecT [81][b]			
Popitam [71][b]			
MS/MS (*De novo*)			
Lutefisk [82, 83][b]			
PepHMM [32][b]			
Sherenga [84][c]			
PepNovo [33][b]			
Peaks (BSI)[c,e]			

Community Development

a) Semi-commercial or must contact author
b) Freely available on the internet
c) Commercial or not available
d) Applied Biosystems
e) Bioinformatics Solutions

Matthiesen, R., *Methods, algorithms and tools in computational proteomics: a practical point of view.* Proteomics, 2007. **7**(16): p. 2815-32.

Software Platforms for Label-free Quantitation

	PNNL Pipeline	PEPPeR	msInspect	SuperHirn	CRAWDAD
Lab	PNNL	Broad Institute	FHCRC	IMSB (Swiss)	Univ. Wash.
Feature Picker	Decon2LS/Viper	Mapquant (or any other)	msInspect	SuperHirn	CRAWDAD
Method	Spectrum de-isotoping then clustering	Image Analysis then de-isotoping	Wavelet decomposition then de-isotoping	Spectrum de-isotoping then merging	m/z channel binning
RT Alignment	Normalization, then linear or LCMSWARP	Relative, then linear, or LOESS (exp)	Iterative non-linear transformation	LOESS modeling	Dynamic time warping
m/z recalibration	Yes (dynamic)	Yes (quadratic)	No	No	No
Assignment of IDs to features	AMT database, normalized elution times	AMT database, relative elution order (Landmarks)	AMT database through user interaction	Yes, but not well documented at present	Yes, for differences only if they exist
Statistical Evaluation of assignment	Mass shift decoy and/or Bayesian Statistics	Bayesian Statistics	No	No	No
Unidentified Feature Recognition	Stored in database for later analysis	Data-dependent tolerance-based clustering	User specified tolerance-based clustering	Tolerance-based merging, heuristics	Difference mapping only
Runs on	Windows with GUI	Web-based (Linux or Windows install bases)	Java with GUI	Linux	Linux/Windows

“Can We Believe These Results ?”

- Credible results require
  - Rigorous statistically models
  - Known FDR or equivalent
  - Validation
    - Measurements
    - Functions
  - Full disclosure of procedures and methods
  - Dissemination
    - Data
    - Custom analysis software tools

- Data standards and release policies are critical
  - DOE GTL data release policy
  - GTL Knowledgebase
  - Martens, L. and H. Hermjakob, *Proteomics data validation: why all must provide data.* Mol Biosyst, 2007. **3**(8): p. 518-22.

Challenges and Opportunities

- Data production rates will increase enormously (e.g. see poster on “Next Generation” Proteomics Platform)
- Data qualities (mass accuracy, CVs, coverage) improving
- The “community problem” is really key for proteomics, since MOST systems share “ill-defined” components to one extent or another (e.g. unintentional adaptive evolution)
- *De novo* approaches increasingly effective and practical, but still demand considerable computational resources

References

- Martens, L. and H. Hermjakob, *Proteomics data validation: why all must provide data.* Mol Biosyst, 2007. **3**(8): p. 518-22.
- Panchaud, A., M. Affolter, P. Moreillon, and M. Kussmann, *Experimental and computational approaches to quantitative proteomics: status quo and outlook.* J Proteomics, 2008. **71**(1): p. 19-33.
- Honda, A., Y. Suzuki, and K. Suzuki, *Review of molecular modification techniques for improved detection of biomolecules by mass spectrometry.* Anal Chim Acta, 2008. **623**(1): p. 1-10.
- Webb-Robertson, B.J. and W.R. Cannon, *Current trends in computational inference from mass spectrometry-based proteomics.* Brief Bioinform, 2007. **8**(5): p. 304-17.
- Matthiesen, R., *Methods, algorithms and tools in computational proteomics: a practical point of view.* Proteomics, 2007. **7**(16): p. 2815-32.
- Bantscheff, M., M. Schirle, G. Sweetman, J. Rick, and B. Kuster, *Quantitative mass spectrometry in proteomics: a critical review.* Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.
- Urfer, W., M. Grzegorczyk, and K. Jung, Statistics for proteomics: a review of tools for analyzing experimental data. Proteomics, 2006. 6 Suppl 2: p. 48-55.

References

- Hernandez, P., M. Muller, and R.D. Appel, *Automated protein identification by tandem mass spectrometry: issues and strategies.* Mass Spectrom Rev, 2006. **25**(2): p. 235-54.
- Peng, J., J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi, *Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome.* J Proteome Res, 2003. **2**(1): p. 43-50.
- Gerber, S.A., J. Rush, O. Stemman, M.W. Kirschner, and S.P. Gygi, *Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.* Proc Natl Acad Sci U S A, 2003. **100**(12): p. 6940-5.
- Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.
- Smith, R.D., G.A. Anderson, M.S. Lipton, L. Pasa-Tolic, Y. Shen, T.P. Conrads, T.D. Veenstra, and H.R. Udseth, *An accurate mass tag strategy for quantitative and high-throughput proteome measurements.* Proteomics, 2002. **2**(5): p. 513-23.
- Smith, R.D., L. Pasa-Tolic, M.S. Lipton, P.K. Jensen, G.A. Anderson, Y. Shen, T.P. Conrads, H.R. Udseth, R. Harkewicz, M.E. Belov, C. Masselon, and T.D. Veenstra, *Rapid quantitative measurements of proteomes by Fourier transform ion cyclotron resonance mass spectrometry.* Electrophoresis, 2001. **22**(9): p. 1652-68.

References

- Conrads, T.P., K. Alving, T.D. Veenstra, M.E. Belov, G.A. Anderson, D.J. Anderson, M.S. Lipton, L. Pasa-Tolic, H.R. Udseth, W.B. Chrisler, B.D. Thrall, and R.D. Smith, *Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling.* Anal Chem, 2001. **73**(9): p. 2132-9.
- Conrads, T.P., G.A. Anderson, T.D. Veenstra, L. Pasa-Tolic, and R.D. Smith, *Utility of accurate mass tags for proteome-wide protein identification.* Anal Chem, 2000. **72**(14): p. 3349-54.
- Nicholson, J.K., J.C. Lindon, and E. Holmes, *'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.* Xenobiotica, 1999. **29**(11): p. 1181-1189.
- Metz, T.O., J.S. Page, E.S. Baker, K.Q. Tang, J. Ding, Y.F. Shen, and R.D. Smith, *High-resolution separations and improved ion production and transmission in metabolomics.* Trac-Trends in Analytical Chemistry, 2008. **27**(3): p. 205-214.
- Metz, T.O., Q. Zhang, J.S. Page, Y. Shen, S.J. Callister, J.M. Jacobs, and R.D. Smith, *The future of liquid chromatography-mass spectrometry (LC-MS) in metabolic profiling and metabolomic studies for biomarker discovery.* Biomark Med, 2007. 1(1): p. 159-185.

Selected Software Resources

- http://omics.pnl.gov (PNNL's Data and Software Distribution Website)
- http://ncrr.pnl.gov (PNNL's NCRR website)
- http://www.sysbep.org/ and http://www.proteomicsresource.org (Salmonella typhimurium data resource)
- http://www.ms-utils.org/ (Magnus Palmblad)
- http://open-ms.sourceforge.net/index.php (European consortium)
- http://tools.proteomecenter.org/SpecArray.php (ISB)
- http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak_Alignment/ (Tobias Kind with Oliver Fiehn)
- http://www.proteomecommons.org/tools.jsp (Phil Andrews and Jayson Falkner)
- http://www.broad.mit.edu/cancer/software/genepattern/ (Broad Institute)
- https://proteomics.fhcrc.org/CPAS/ (FHCRC)