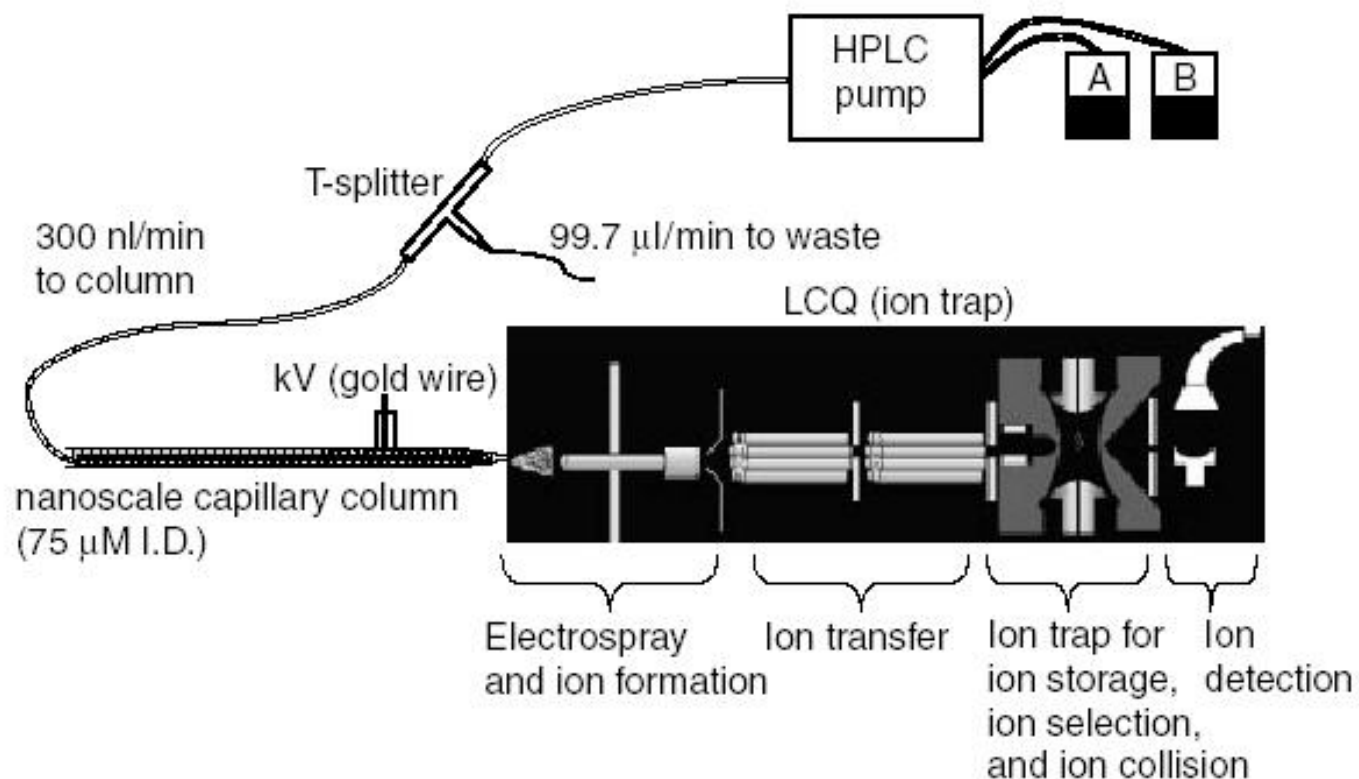


GenomicPeptideFinder:

Mass spectrometric data mining in genomic data bases

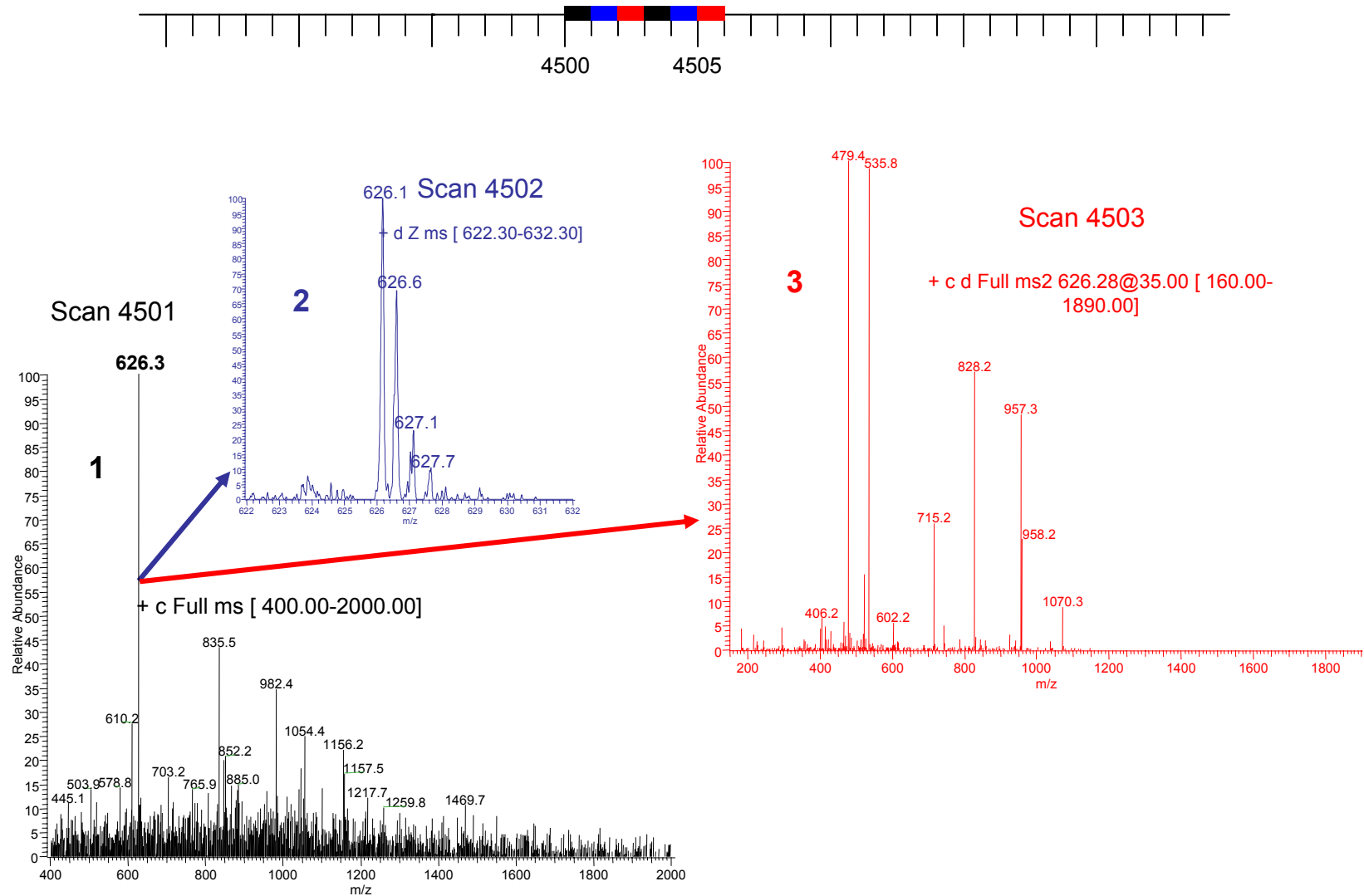
Institute for Plant Biochemistry and Biotechnology
University of Münster, Germany

Schematic depiction of an ion trap mass spectrometer

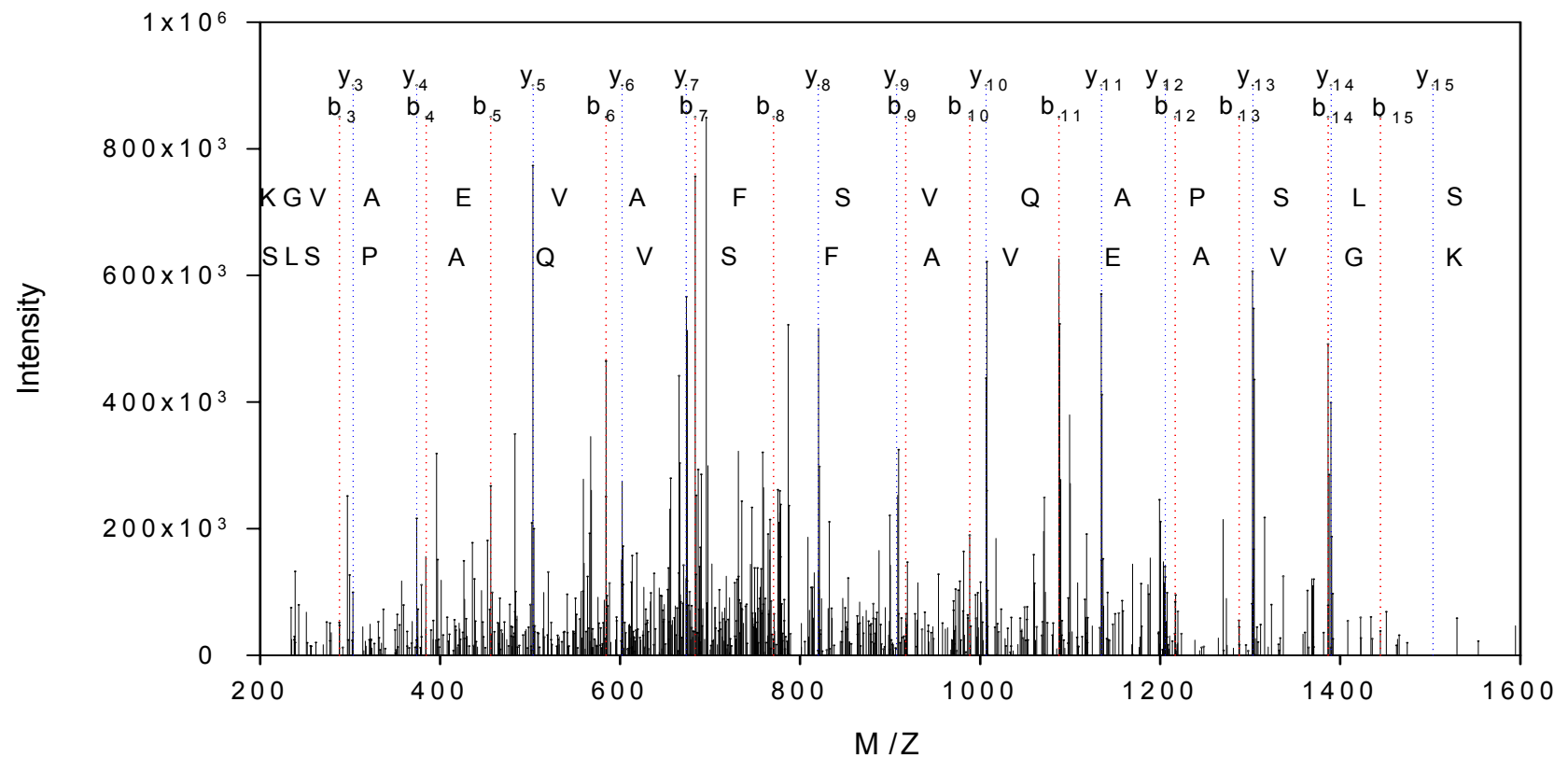
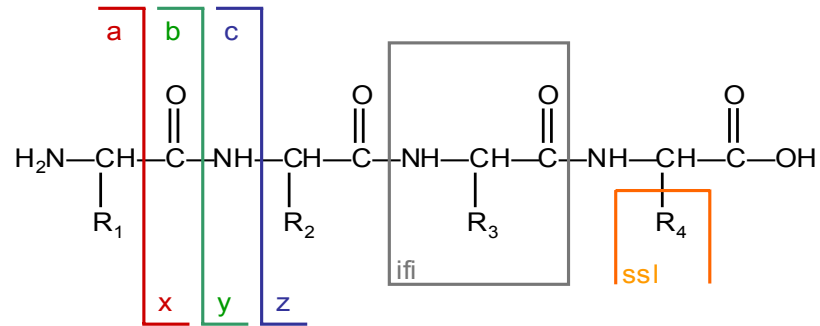


Peng, J. and Gygi, S.P. (2001) Proteomics: the move to mixtures. *J. Mass Spectrom.*, **36**, 1083-1091.

Mass Spectra Recording (Triple Play)



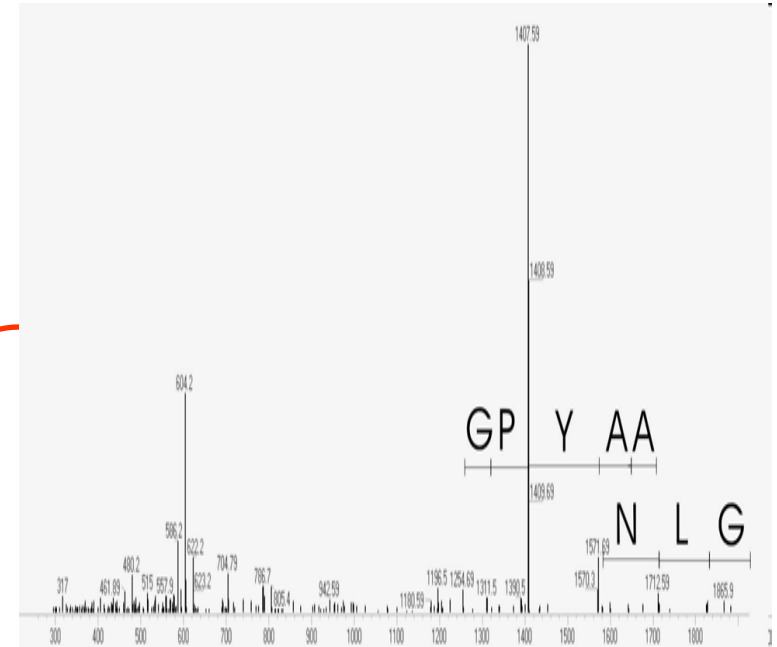
Fragmentation Spectrum



Correlation analysis

Database selection

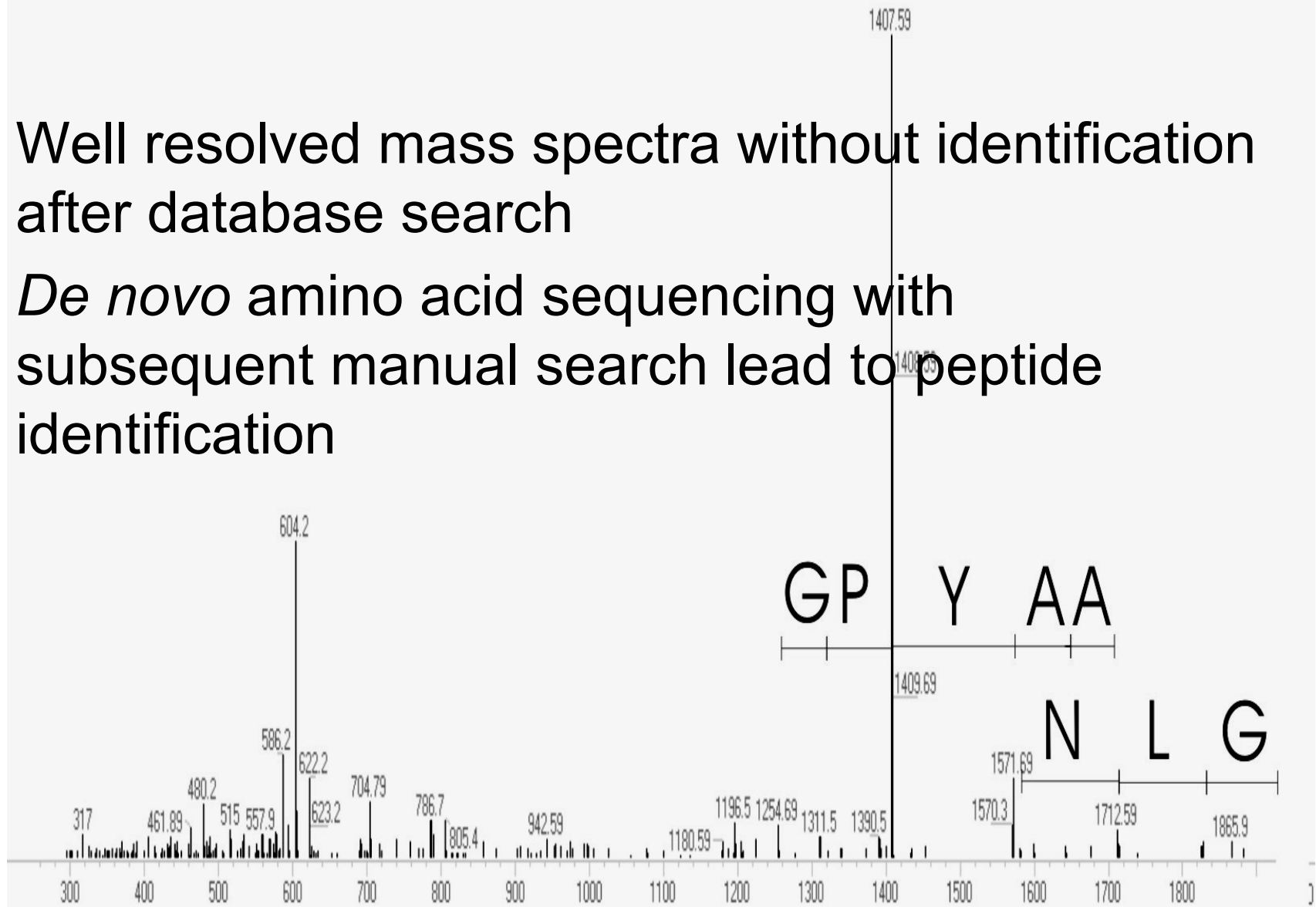
>1080ZR
 IAAYPGVSPGLMIHYNIGR
 >1137RZ
 AAYPGATQPGATELARRLGK
 >1152RZ
 GSGDAAYPGGPFFNLFNLGK
 >1152ZR
 GSGDAAYPGGPFFNLFNLGK
 >2360RZ
 VDSGWGGVVVALAPYNLGR
 >240RZ
 HPGVVCPRPGRGGGCSRHIK
 HPGVVCCSRHRRSHTIGK



#	deltCn	XCorr	Sp	Reference	Peptide
1.	0.0000	4.2902	382.4	1152RZ	(-)GSGDAAYPGGPFFNLFNLGK
2.	0.5359	1.3963	70.0	582RZ	(V)VCLMSAIALPYNGGLRPPGV
3.	0.6028	1.1949	43.1	258RZ	(-)GGGVVCFDASFRSPKFGIGK
5.	0.6160	1.1553	70.1	2360RZ	(-)VDSGWGGVVVALAPYNLGR
6.	0.6666	1.0029	37824	37RZ	(K)CVLWAYPGVALGRVQVLGK
11.	0.7333	0.8022	37647	304RZ	(V)GGWLGRVLGGATVLGKAYPGV
12.	0.7351	0.7970	29.2	1080RZ	(-)IAAYPGVSPGLMIHYNIGR
13.	0.7774	0.6695	22.0	304RZ	(K)AYPGVGELVRGGCMMCSIGK
14.	0.7842	0.6492	31.4	37RZ	(W)AYPGVALGRGIRLGKCVLW

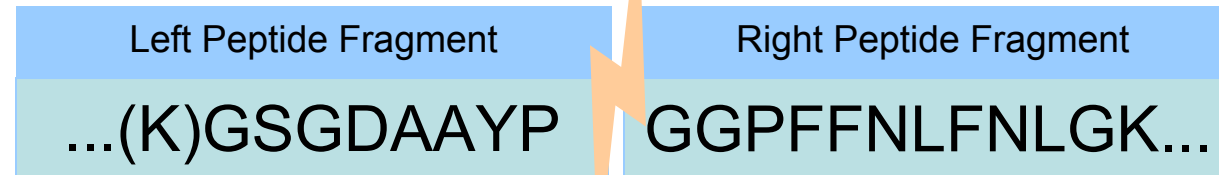
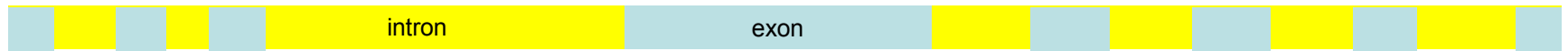
Problem

- Well resolved mass spectra without identification after database search
- *De novo* amino acid sequencing with subsequent manual search lead to peptide identification



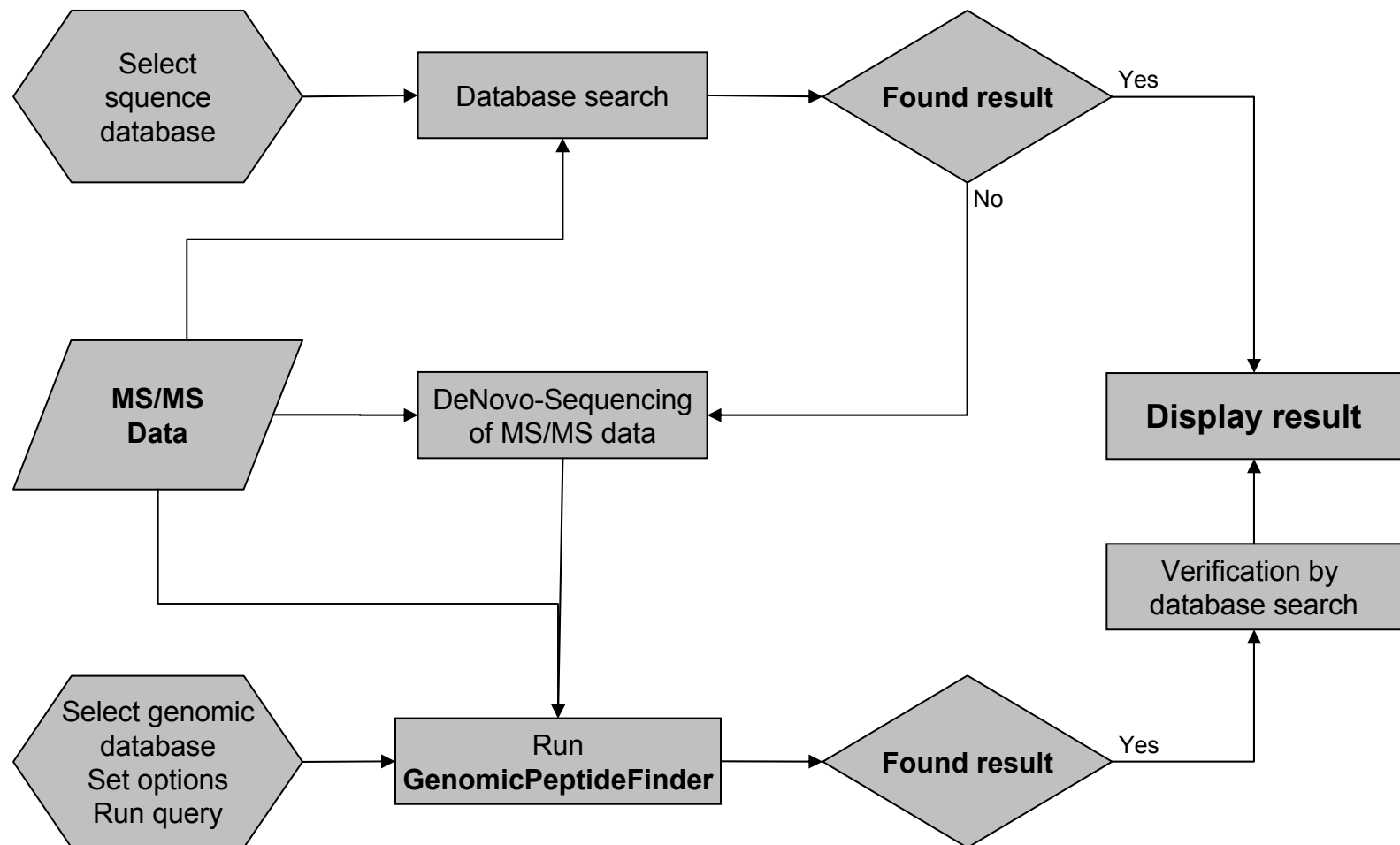
Sequence availability

- Every fourth peptide escapes detection by MS due to an intron which splits the peptide on the “genomic level”
 - Choudhary, J. S., Blackstock, W. P., Creasy, D. M. and Cottrell, J. S. (2001) *Interrogating the human genome using uninterpreted mass spectrometry data. Trends Biotechnol*, **19**, S17-22.
- Each human and Chlamy gene statistically contains 8 exons
 - Deutsch, M. and Long, M. (1999) *Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res*, **27**, 3219-3228.
 - Merchant, S. et al., (2007) *Science*; in press
- 40-60% of human genes are alternatively spliced
 - Modrek, B. and Lee, C. (2002) *A genomic view of alternative splicing. Nat Genet*, **30**, 13-19.
- 50% of a genome may not be complementable by EST data
 - Seki, M., Narusaka, M., Kamiya, A., Ishida, et. al. (2002) *Functional annotation of a full-length Arabidopsis cDNA collection. Science*, **296**, 141-145.
- Gene prediction is far from being correct (Only 15-20% of known gene structures are predicted correctly.)
 - Brent, M.R. and Guigo, R. (2004) *Recent advances in gene structure prediction. Curr Opin Struct Biol*, **14**, 264-272.
- Post translational modifications
 - Perkins, D. N., Pappin, D. J., Creasy, D. M. and Cottrell, J. S. (1999) *Probability-based protein identification by searching sequence databases using mass spectrometric data. Electrophoresis*, **20**, 3551-67.

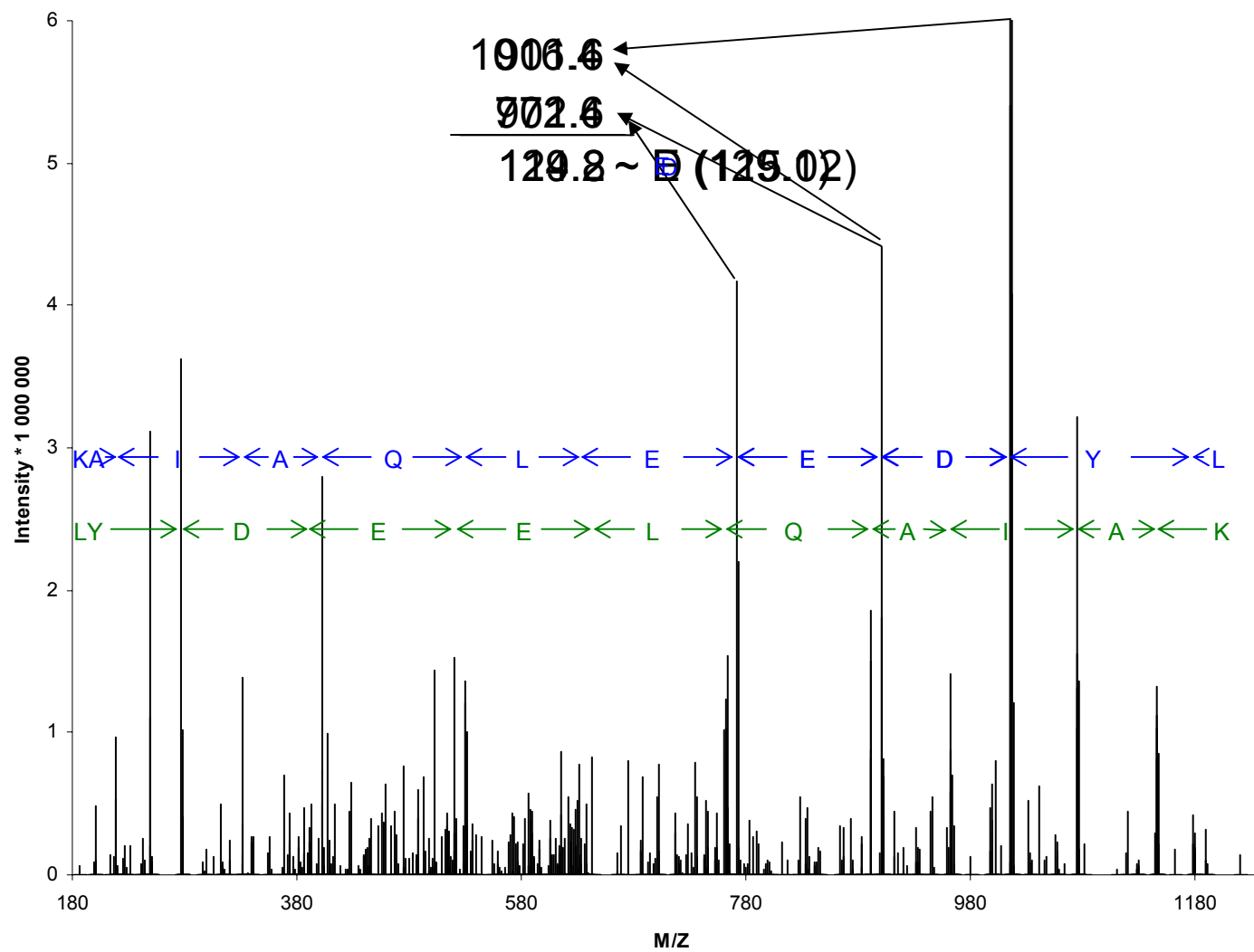


GTGAGCAGATAGAGGAGAGAGCGCGCGAGAGAGGGCCGCGGC
TGCAGCTGGTGTGGGCCAAGGCTGGCGGAGACAGCCAGCAGG
GGTCAGGGGGAGGGGACAGGGCAGAGGCAGACGCCGGCGCT
GTCGGGTGCGGGATGCGGGTTGCTGATGGATGGAAACTCCGA
TCCGCGGCGGACGGTTTCGTGCTATGTAGCTGCTCAACAGGG
TTTGCATGCTCCTGGCTGACAGCGCATATGACGGTCCCTTCC
CCCGCTCCGCAG

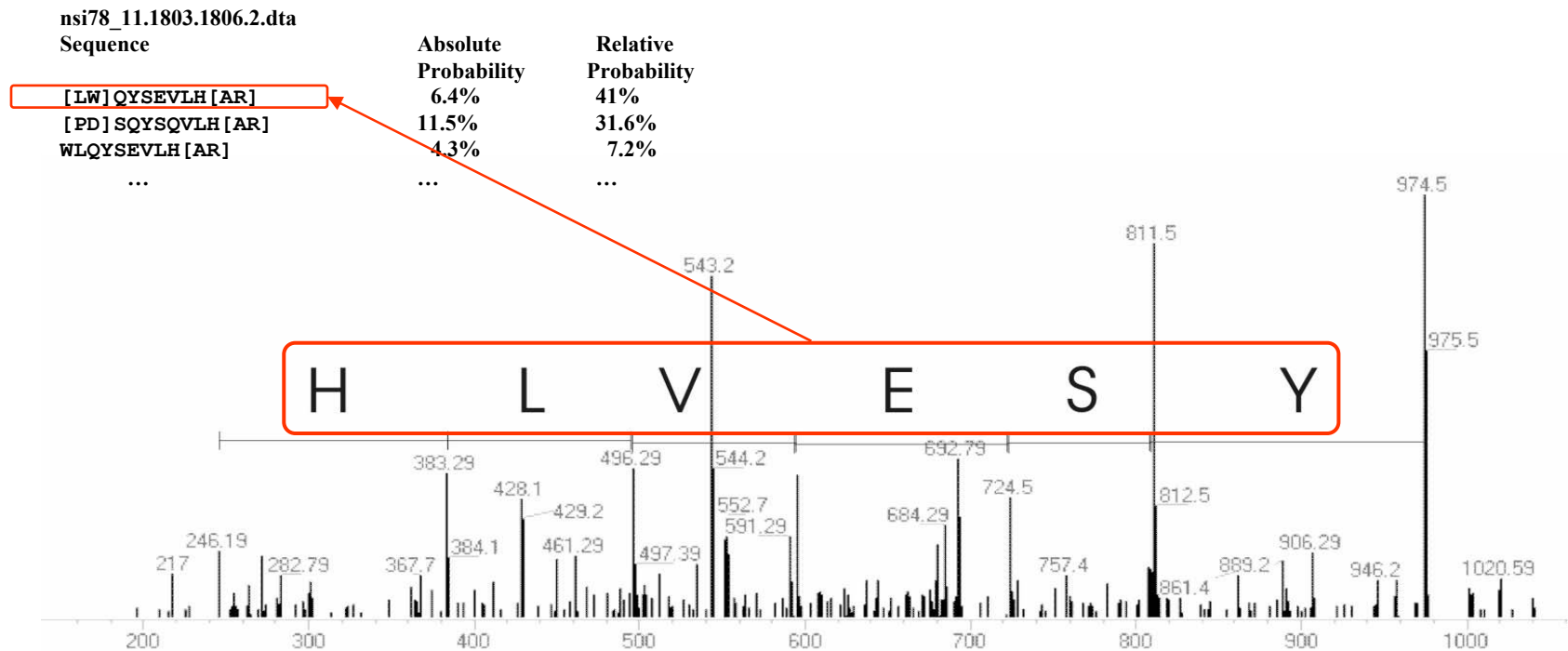
Intron

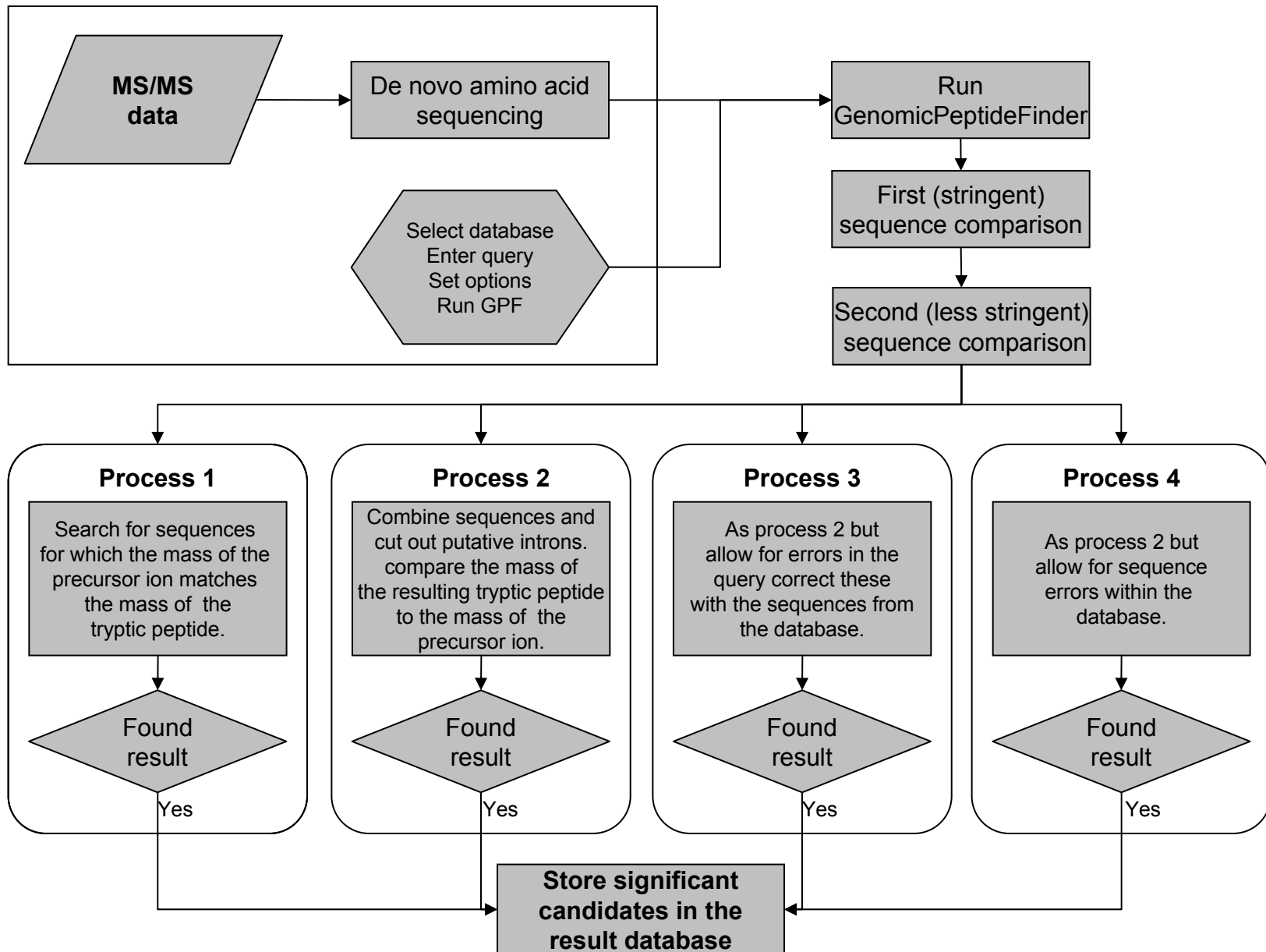


De Novo Sequencing



De Novo Amino Acid Sequence Predictions





WLQYSEVIHAR

1402.4 Da

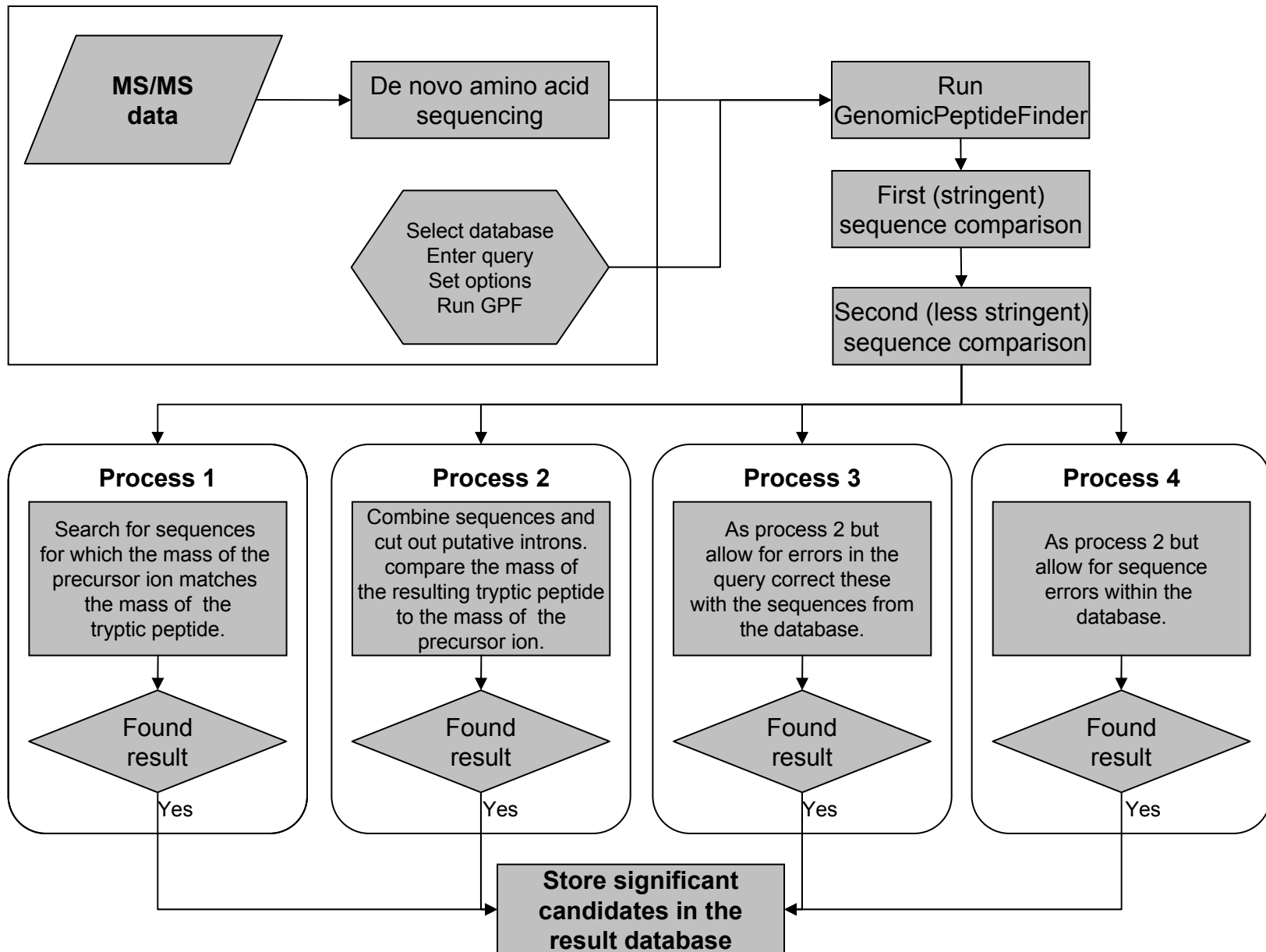
nsi_63_1.1641.1644.2

RF1: PDPRCHQHQQVVRERRNPPRRLVQRLLG-PLHHLLRGDRGYAVR-AAPPSGAGREGQALVG
RF2: SIFRHPKQDFGFDPLGLLDPVNSGGFIEPKWLQYSEVIHARWAMLGAAGCIAPEVLGAAG
RF3: -SPMPPTSSGSRAA-SPPPARTTATGLTPTPSSSWRSWLCSSLSCAAFCGEGGAGARRA
RF4: PLLAVSASLGPHQLQPRPSLARSLLYLLTGVGSVARALEDRLQAQEVLLAHGAGVAEVLQ
RF5: PCWLSPPALAHTSCSRGPLSRALSSICSPG-AASPEPLKIASRPRKYCWPMEPG-RKSCR
RF6: THARRGSVIGCNCNQPQLQRCTRGEK-NRFLRRFFFFARRAPAPPSPHLKAAQLSELHSH

Measured Mass: **1402.40 Da**

Calculated Mass: **1401.58 Da**

Mass Difference: **0.82 Da** < 700 ppm (0.98 Da)



[RC*] AAYPG[VV]CFNPYNLGK

2027.95

nsi78_03.2460.2463.2

RF1: DFRYPGSMGQQYFLGLEAIFKGS~~GD~~AAYPG~~EO~~EERARERGPRILQLVWAKAGGDSQQGSGGGAQGRGRRRRRCRVRDAGC
RF2: TSATPAPWASSTSWAWRRSSRALATLPVSR_RRERAREGRGCSWCGPRLAETASRGQGEHRAEADAGAVGCGMRVAD
RF3: LPLPRLHGPAVLPGPGGDLQGLWRRCLPR_ADRGESARERAAAAAGVGQGWRRQPAGVRGRGTGQRQTPALSGAGCGLL
RF4: QAAVLDDLLELQLLHGSGLGLAQVEQVEEGAACGAGEGTVICAVSQEHANPVEQLHSTKPSAADRSFHPSATRIHPHTAPASAS
RF5: RRPFLISLSFSSFMAASVLPRLNRLKKGPPAERGKGPSYALSARSMQTLSSYIARNRPPRIGVSIHQPPASRTRQRRRLPLPCA
RF6: GGRS_SP_ASAPSWQPRSCPG_TG_RRGRLRSGGRDRHMRCQPGACKPC_AAT_HETVRRGSEFSPISNPHPAPDSAGVCLCPV

~~GS~~GD AAYPG~~EO~~EERENLGK

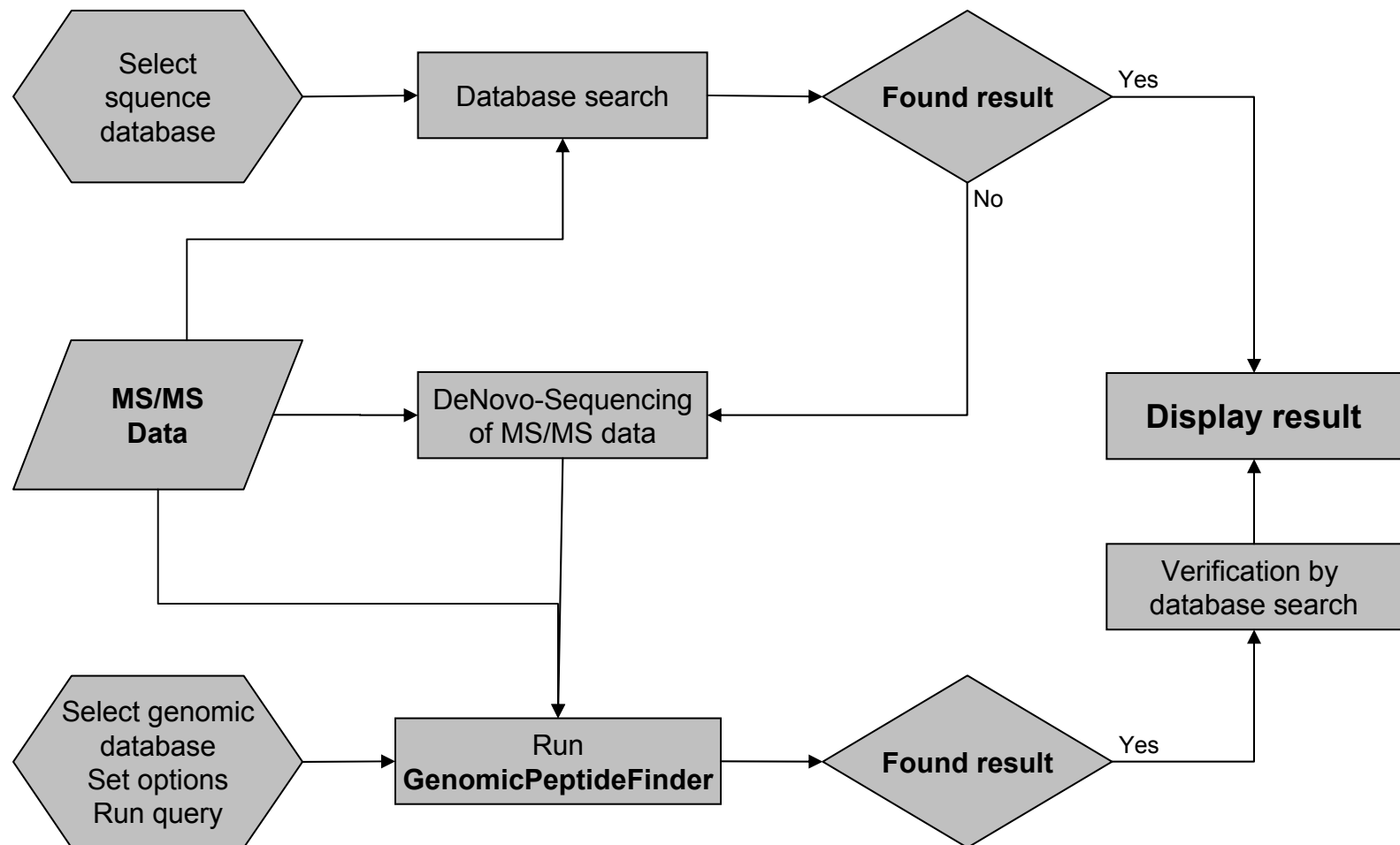
Measured Mass ~ Calculated Mass

RF1: WMETPIRGGRFRAM_LLNRVCMLLADSAYDGPFP~~RS~~AGGPF~~FN~~LNLEKTEAAMKELKLKEIKNGRL
RF2: GWKLRSAADGFVLCSCSTGFACSWLTAHMTVPSPAPQAAPSS~~TC~~STWARPLP_RS_SSRRSRTAA
RF3: MDGNSDPRRTVSCYVAAQQGLHAPG_QRI_RSLPPLRRRPLLQPVQPGQDRGCHEGAEAQGDQERPP
RF4: ALCPSPLLA~~V~~SASLGP~~H~~QLQPRPSLARSLLYLLTG~~V~~GSVARALEDRLQAQEVLLAHGAGVAEV
RF5: PLPLTPAGCLRQPWPTPAAAAALS~~R~~ALSPLSAHRGRQRRQSP_RSPPGPGSTAGPWSRSGSGS
RF6: DLRVLQPLGLDEAAAVHGVQQAQGV~~E~~AEVLFGMTKNGQ-IALMLDRGGMSGHADFI~~A~~ERL

Measured Mass: 2027.95 Da

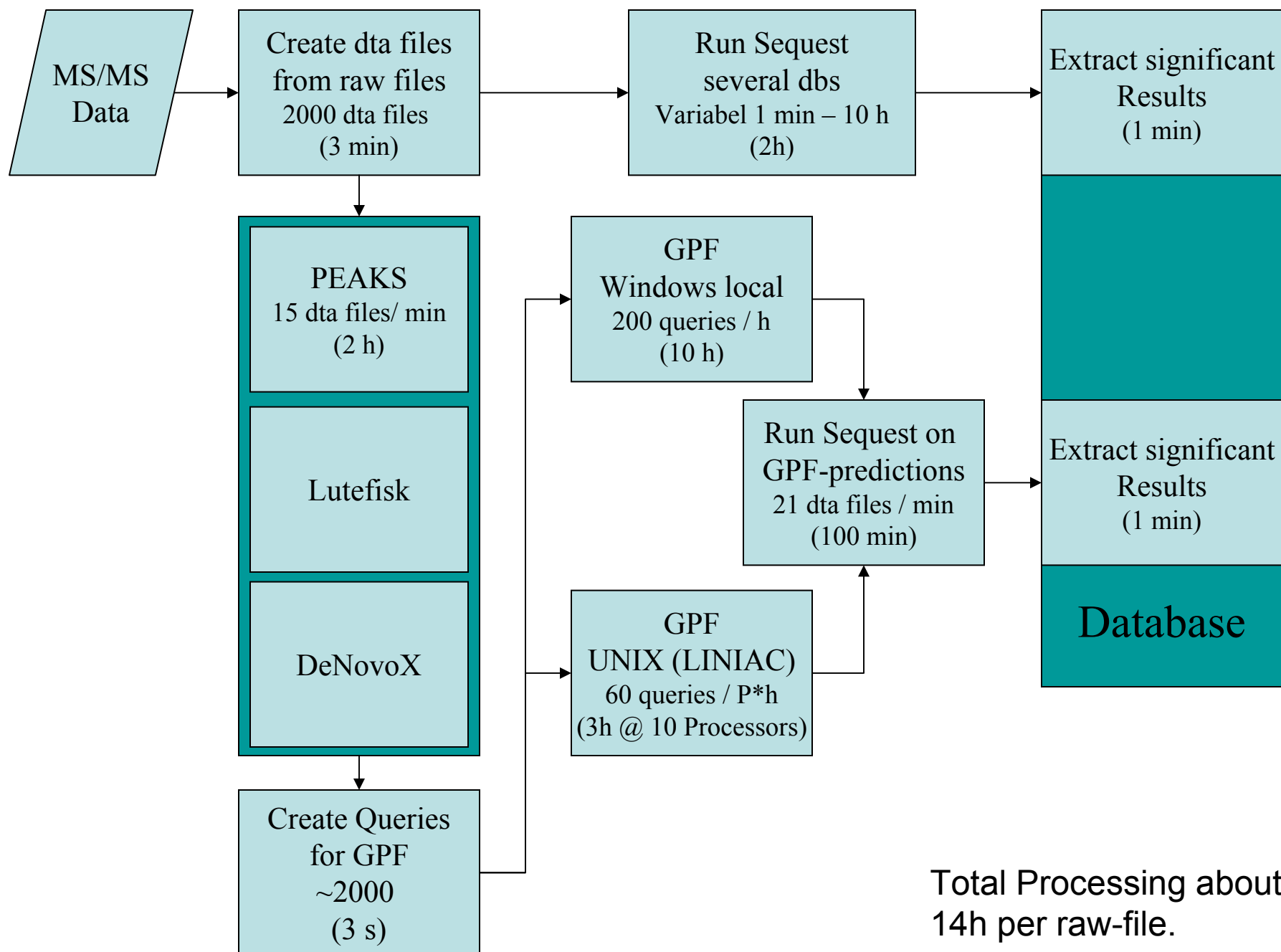
Calculated Mass: 2027.98 Da

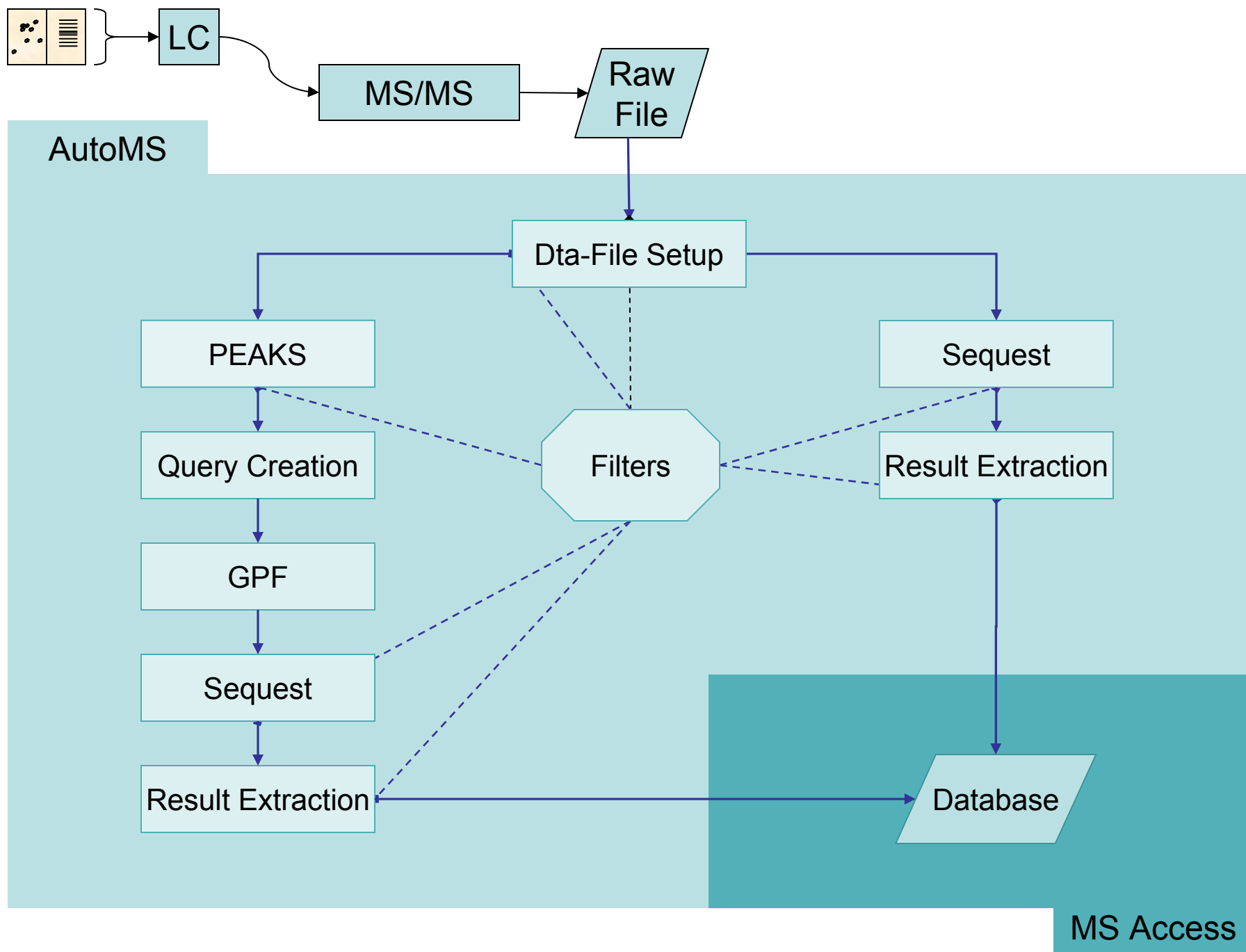
Mass Difference: 0.03 Da < 700 ppm (~ 1.4 Da)



Analysis Workflow: Analysis of thylakoid membranes after 1D-SDS-PAGE







Identified Peptides

Confident peptide identifications and complementarities

	Sequest	GPF	PEAKS	Complement
Sequest	2496(22)	322 (19)	72 (5)	
GPF	322 (19)	448 (98)	60 (17)	
PEAKS	72 (5)	60 (17)	105 (17)	
Complement				47 (4)

Number of identified peptides (non redundant) is 2622

Complement PEAKS-Sequest: 3%
Improved by using GPF : 18%

Peptides	Found in gene models	missed	%-missed
Sequest	1317	1157	47
GPF	332	58	15
Comple- ment	292	11	4

Example 1 – Gene model verification

Peptides

ALE-GEIYDTFK, bnTY3_06302004.1322.1322.2.dta (Xcorr = 4.1)
MEGEFTADNVAK, bnTY3_06302004.920.922.2.dta (Xcorr = 4.12)
TGPPAVTVEDADK, bnTY3_06302004.742.746.2.dta (Xcorr = 3.26)
TPIVFEGGDR, bnTY3_06302004.1003.1003.2.dta (Xcorr = 2.84)

Scaff: 177403 SDGIVGWVKKKTGPPAVTVEDADKLKSLEADAEV VVVGYFK**ALE** 177272
 +DGIVGWVKKKTGPPAVTVEDADKLKSLEADAEV VVVGYFKALE
 sbjct: 145 ADGIVGWVKKKTGPPAVTVEDADKLKSLEADAEV VVVGYFK**ALE** 188

Scaff: 176994 Q**GEIYDTFK**SYAAKTEDVVFVQTTSDAVAKAAGLDAVDTVSVVKNFAGE 176848
 +GEIYDTFKSYAAKTEDVVFVQTTSDAVAKAAGLDAVDTVSVVKNFAGE
 sbjct: 188 E**GEIYDTFK**SYAAKTEDVVFVQTTSDAVAKAAGLDAVDTVSVVKNFAGE 236

**JGI: tblastn
 blast protein
 vs.
 translated nucleotide**

Gene model: C_390061 [chlre2:164450]

MNRWNLLALTLG LLLVAAPFTKHQFAHASDEYEDDEEDDAPAAPKDDVDVTVVTVKNWDETVKKS KFALVE
 FYAPWCGHCKTLKPEYAKAATALKAAAPDALIAKVDATQEE SLAQKFGVQGYPTLKWFDGELASDYNGPRD
 ADGIVGWVKKKT**TGPPAVTVEDADKLKSLEADAEV VVVGYFKALEGEIYDTFK**SYAAKTEDVVFVQTTSDAVAKAAGLDAVDTVSVVKNFAGEDRATAVLATDIDTDSLTA FVKSEKMPPTIEFNQKNSDKIFNSGINKQL
 ILWTTADDLKADAEIMTVFREASKKFKGQLVFVTVNNEGDGADPVTNFFGLKGATSPVLLGFFMEKNKKFRM
EGEFTADNVAKFAESVVDGTAQAVLKSEAIPEDPYEDGVYKIVGKTVESVVLDETKDVLLEVYAPWCGHCKK
 LEPIYKKLAKRFKKVDSVIIAKMDGTENEHPEIEVKGFPTILFY PAGSDR**TPIVFEGGDR**SLKSLTKFIKTN
 AKIPYELPKKGS DGDEGTSDDKDKPASDKDEL*

JGI annotation: Protein disulfide isomerase 1 (CrPDI1), encodes the chloroplastic Rb60 protein which is part of a complex involved in the control of psbA mRNA translation by light

Scores

NCBI: 0
[gi|4104541|gb|AAD02069.1](#)
 protein disulfide isomerase
 SIB-BLAST: 0
 _CHLRE Protein disulfide isomerase RB60 (EC 5.3.4.1)
 Kazusa: 0
[AF036939.1](#) Chlamydomonas reinhardtii protein disulfide isomerase
 Other:

1

Intron split and parts of intron split peptides are in red: **ALE-GEIYDTFK**
 Significant non-intron split peptides are in blue: **MEGEFTADNVAK**

Example 2 - Gene model correction

Peptides

LEEGE-MPLNTYSNK, bnTY5a_062904.1079.1083.2.dta (**Xcorr = 4.17**)

LYDEEL-QAIAK, bnTY5a_062904.1218.1222.2.dta (**Xcorr = 4.14**)

1

Gene model: C_1410005[chlre2:155574]

Scores

MQTVRAPAASGVATRVAGRRMCRPVAATKASTAVTTDMSKRTVP**TKLEEGE**VSACLGATTGRRKMGRGGE
RRSGSVGQELAGRILLICDFLFFILQ**MPLNTYSNK**APFKAKVRSVEKITGPKATGETCHII IETEGKIP
FWEGQSYGVIPPGTKINSKGKEVPHGTRLYSIASSRYGDDFDGQTASLCVRRAVYVDPETGKEDPAKKGLCS
NFLCDATPGTEISMTGPTGKVLLLPADANAPLICVATGTGIAPFRSFWRRCFIENVPSYKFTGLFWLFMGVA
NSDAK**LYDEELQAIAK**AYPGQFRLDYALSREQNNRKGKMYIQDKVEEYADEIFDLLDNGAHMYFCGL
KGMMPGIQDMLERGLERNSGESSEELLRSRTRQGAGGGGTRTRRAAGRAGGMPRRKAKAPRAGRGTSMMSG
AV*

JGI annotation: Ferredoxin--NADP reductase, chloroplast precursor. The model doesn't completely fit the published cDNA sequence gi|1706781|sp|P53991|FENR_CHLRE[1706781],

NCBI: 9e-162

gi|732532|gb|AAB40978.1
ferredoxin-NADP+ reductase

SIB-BLAST: 1e-162

_VOLCA Ferredoxin-NADP+
reductase [fnr] [Volvox carteri]

Kazusa: 1e-170

U10545.1 Chlamydomonas
reinhardtii ferredoxin-NADP+
reductase

Other:

2

cDNA sequence: **FNR cDNA /FENR_CHLRE**

Scores

MASLRKPSNHADRACSRRLRVATRVAGRRMCRPVAATKASTAVTTDMSKRTVP**TKLEEGE****MPLNTYSN**
KAPFKAKVRSVEKITGPKATGETCHII IETEGKIPFWEGQSYGVIPPGTKINSKGKEVPTARLYSIASSRYG
DDGDGQTASLCVRRAVYVDPETGKEDPAKKGLCSNFLCDATPGTEISMTGPTGKVLLLPADANAPLICVATG
TGIAPFRSFWRRCFIENVPSYKFTGLFWLFMGVNSDAK**LYDEELQAIAK**AYPGQFRLDYALSREQNN
RKGKMYIQDKVEEYADEIFDLLDNGAHMYFCGLKGMMPGIQDMLERVAK
EKGLNYEEWVEGLKHKNNQWHVEVY

[gi|1706781|sp|P53991|FENR_CHLRE\[1706781\]](http://gi|1706781|sp|P53991|FENR_CHLRE[1706781])

NCBI: 0

gi|1706781|sp|P53991|FENR_CHLRE Ferredoxin NADP

reductase

SIB-BLAST: 0

FENR_CHLRE Ferredoxin--
NADP reductase, chloroplast

Kazusa: 0

U10545.1 Chlamydomonas
reinhardtii ferredoxin-NADP+
reductase

Other:

Example 3 – Alternative splicing?

Peptides

VNGGPAGEGLDPL-ADDPDTFAELK, bntyl27a_08182004.2232.2236.2.dta, Xcorr 3.64

1

Gene model: C_1460037 [chlre2:155818]

MAFALATSRKALQVTCKATGKKTAAKARIISYNQIEIAAAPKSSGVEFYGPNRAKWLGPYSENSTPAYLTGE
FPGDYGWDTAGLSADPETFKRYRELELIHARWAMLGALGCLTPELLAKSGTKFGEAVWFKAGAQIFSEGGLD
YLGNP SLVHAQNIVATLAVQVILMGLIEGYR**VNGGPAGEGLDPL**YPGESFDPLGL**ADDPDTFAE**
LKVKEIKNGRLAMFSMFGFFVQAIVTGKGPVQNLDDHLANPGVNNAFAFATKFTPSA

JGI annotation: Lhcbm4

Scores

NCBI: 2e-136

[gi|4139216|gb|AAD03731.1|](#)
light harvesting complex II

SIB-BLAST: 1e-136

tr [Q9ZSJ5](#) CHLRE [Lhcb2]

Kazusa: -

2

Gene model: C_1460037 [chlre2:155818]

MAFALATSRKALQVTCKATGKKTAAKARIISYNQIEIAAAPKSSGVEFYGPNRAKWLGPYSENSTPAYLTGE
FPGDYGWDTAGLSADPETFKRYRELELIHARWAMLGALGCLTPELLAKSGTKFGEAVWFKAGAQIFSEGGLD
YLGNP SLVHAQNIVATLAVQVILMGLIEGYR**VNGGPAGEGLDPLADDPDTFAELK**VKEIKNGRL
AMFSMFGFFVQAIVTGKGPVQNLDDHLANPGVNNAFAFATKFTPSA

Altered model is not supported by EST data

Scores

NCBI: 1e-126

[gi|4139216|gb|AAD03731.1|](#)
light harvesting complex II

SIB-BLAST: 1e-126

tr [Q9ZSJ5](#) _CHLRE [Lhcb2]

Kazusa: -

GenomicPeptideFinder

- Detection of peptides in genomic databases
 - Increases yield 10 fold when compared to de novo sequencing alone
- Detection of peptides split by introns when deduced from genomic DNA
 - Enables identification of approximately 20% formerly undetectable spectra
- Gene Annotation
 - Enables gene annotation from proteomic data
- Additionally
 - Detection of proteins by a single supporting peptide significantly enhances identification of
 - Short proteins
 - Low abundant proteins
 - Modified proteins
 - Alternatively spliced proteins
 - High-Throughput possibility

Allmer et al. (2004) FEBS Lett 562, 202-6.

Allmer et al. (2006) Proteomics 6, 6207-6220.

Third generation GPF

- GPF 2: currently in development
- improvements:
 - also finds intron splits within nucleotide triplets
 - higher speed
 - ~100 times faster than the old GPF
 - 3 queries per second on average
 - speedup due to the fact that an index is used and thus peptides can be “looked up” instead of “searched for”
 - includes visualization of hits
- could be useful for genome annotation

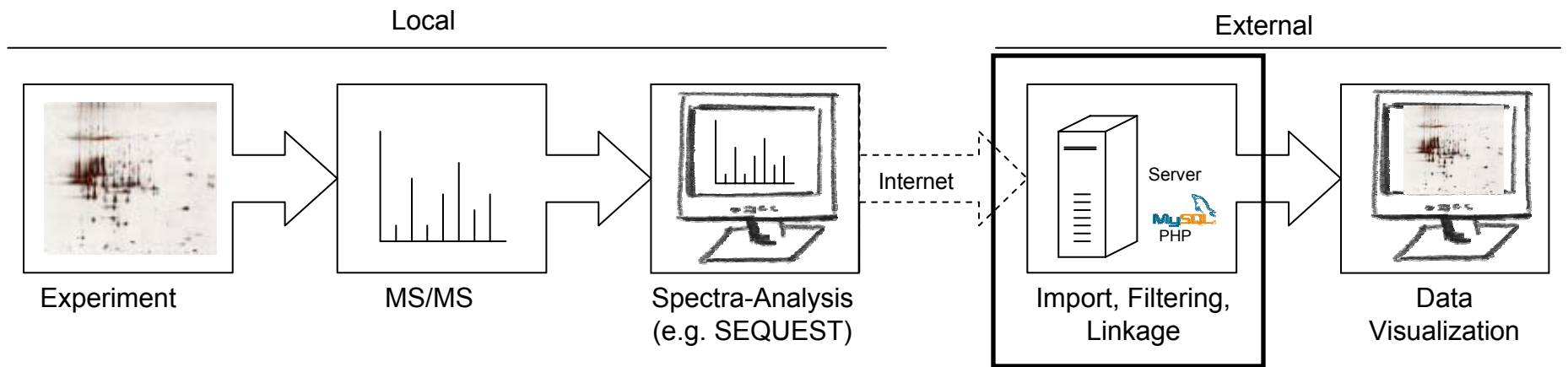
Third generation GPF

- a web interface can be found at www.uni-muenster.de/hippler/gpf
- use **gpf** as user name and **hellofellow** as password for the authentication
- GPF 2 is still work in progress, so the service may become temporarily unavailable or move to another location

2DB (relational database)

- Replaces PepProtDB
 - Web based
 - Multiuser
 - Fully searchable
 - Enhanced visualization

Workflow

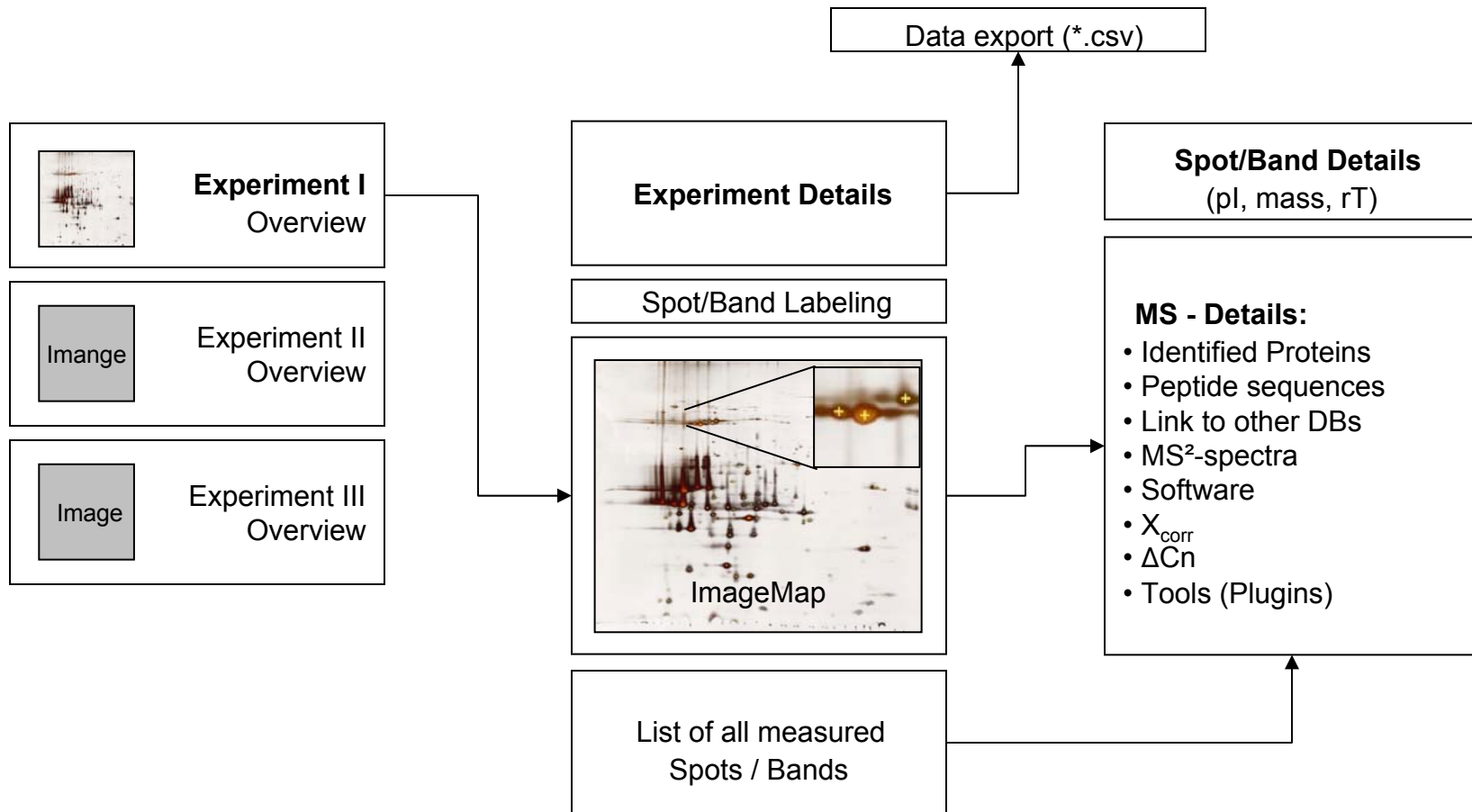


Application Structure

Experiment Overview

Experiment Details

Spot/Band Details



Conclusion

- 2DB speeds up data analysis
 - Automates protein identification
 - Provides results in Tables
 - Data organisation facilitates analysis

Acknowledgement



FRIEDRICH SCHILLER
UNIVERSITÄT
JENA

Einar J. Stauber

Blandine Rimbault



Frederik Sommer

Christine Markert

Monica Zhang

Joshua Goldenberg



WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER

**Institute of Plant Biochemistry
and Biotechnology**

Jens Allmer

Andreas Busch

Marita Hermann

Susan Hawat

Bernadeta Kukuczka

Sebastian Kuhlert

Bianca Naumann

Elisabeth Ostendorf

Michael Specht

Mia Terashima

Michael Hippler

Cooperations

David Roos
UPENN, USA

Funding

Deutsche Forschungsgemeinschaft
Research Grant UPENN