Proteomics data

repositories: PRIDE

Dr. Juan Antonio VIZCAINO

PRIDE Group coordinator

PRIDE team, Proteomics Services Group
PANDA group
European Bioinformatics Institute
Hinxton, Cambridge
United Kingdom

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

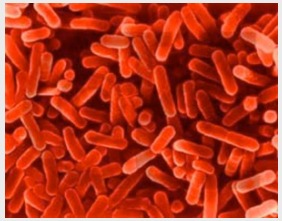
Overview …

- Why sharing proteomics data?
- Introduction to existing proteomics repositories
- Proteomics data bottlenecks
- PRIDE in detail…
- ProteomeXchange consortium

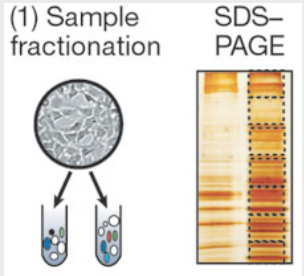
A ONE-SLIDE INTRODUCTION TO MASS SPEC PROTEOMICS

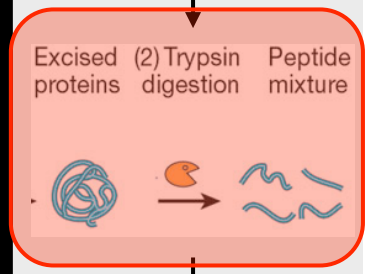
Juan A. Vizcaíno
juan@ebi.ac.uk

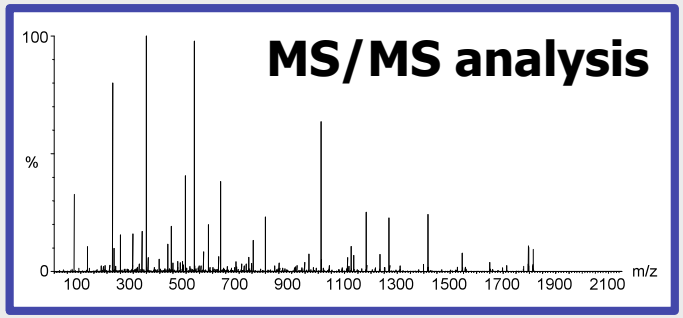
MS proteomics: overall workflow



PROTOCOL





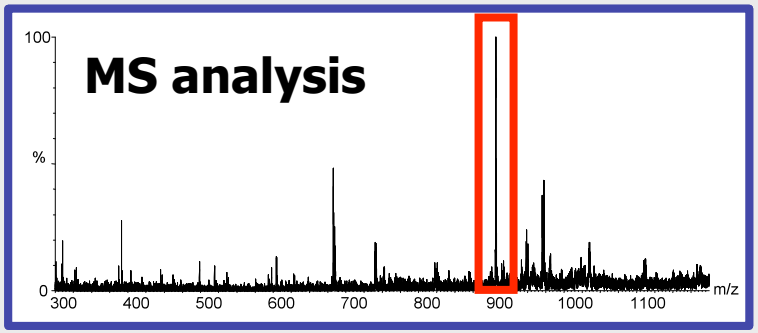


peptides

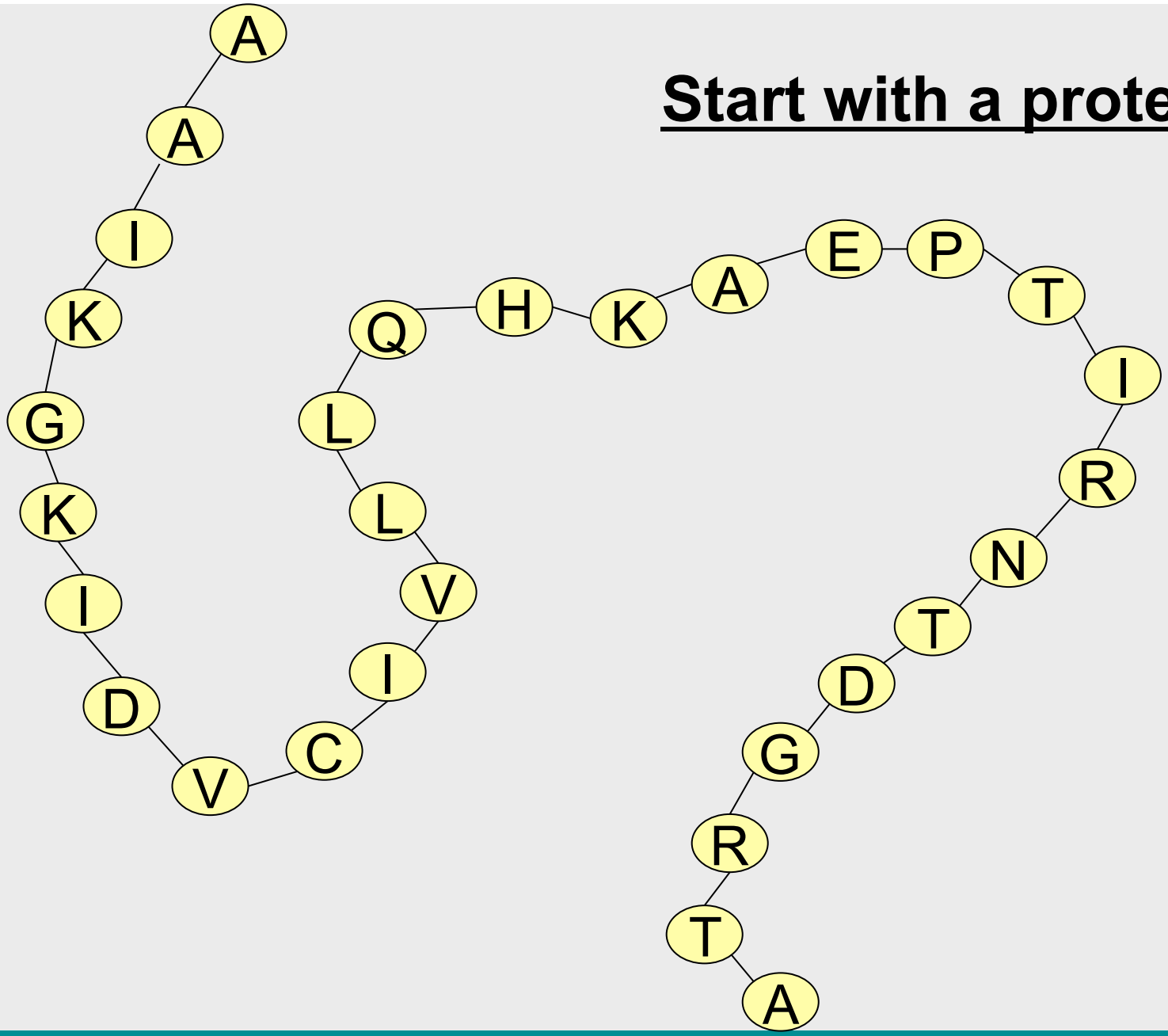
sequence database

proteins

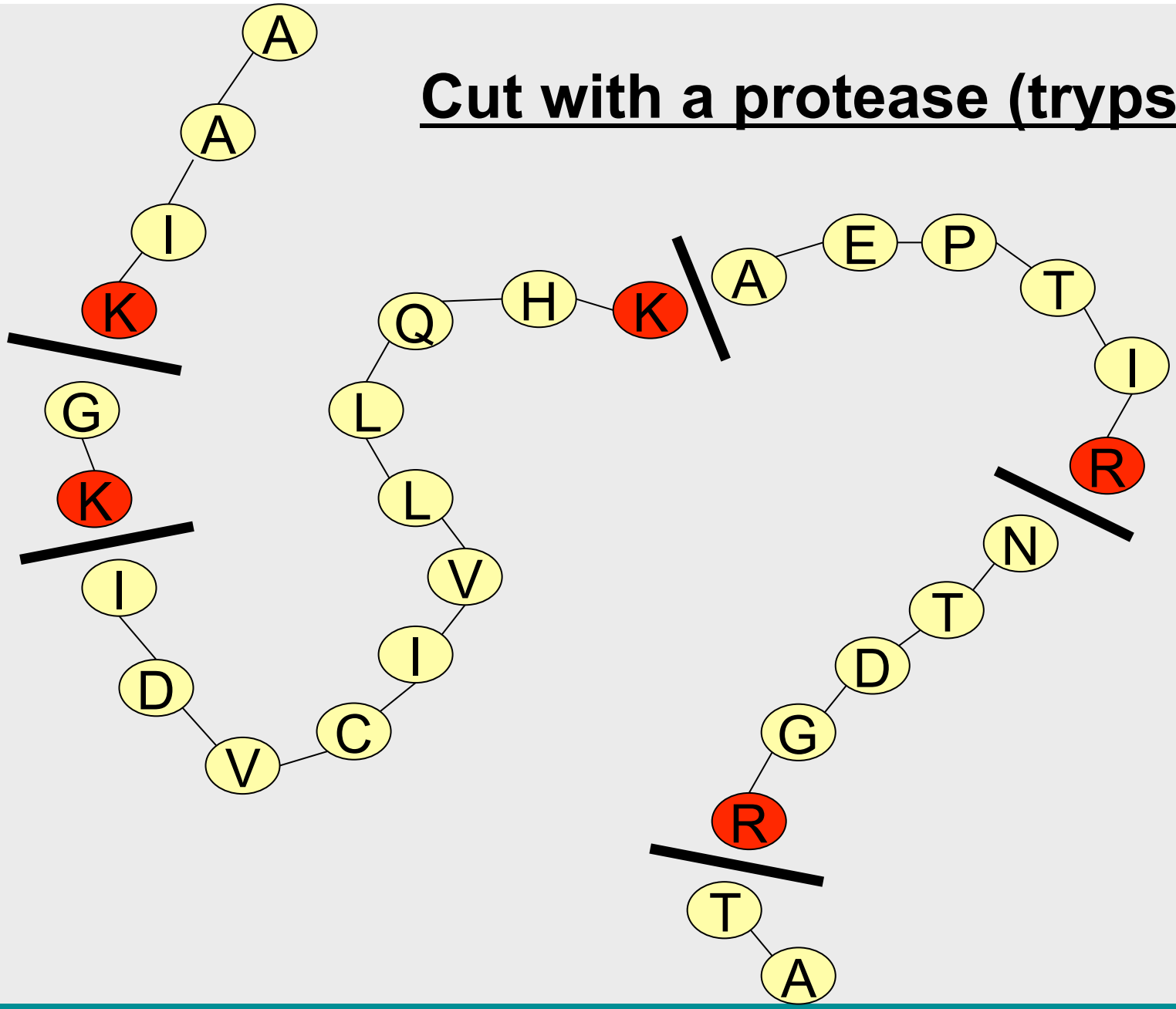
fragmentation



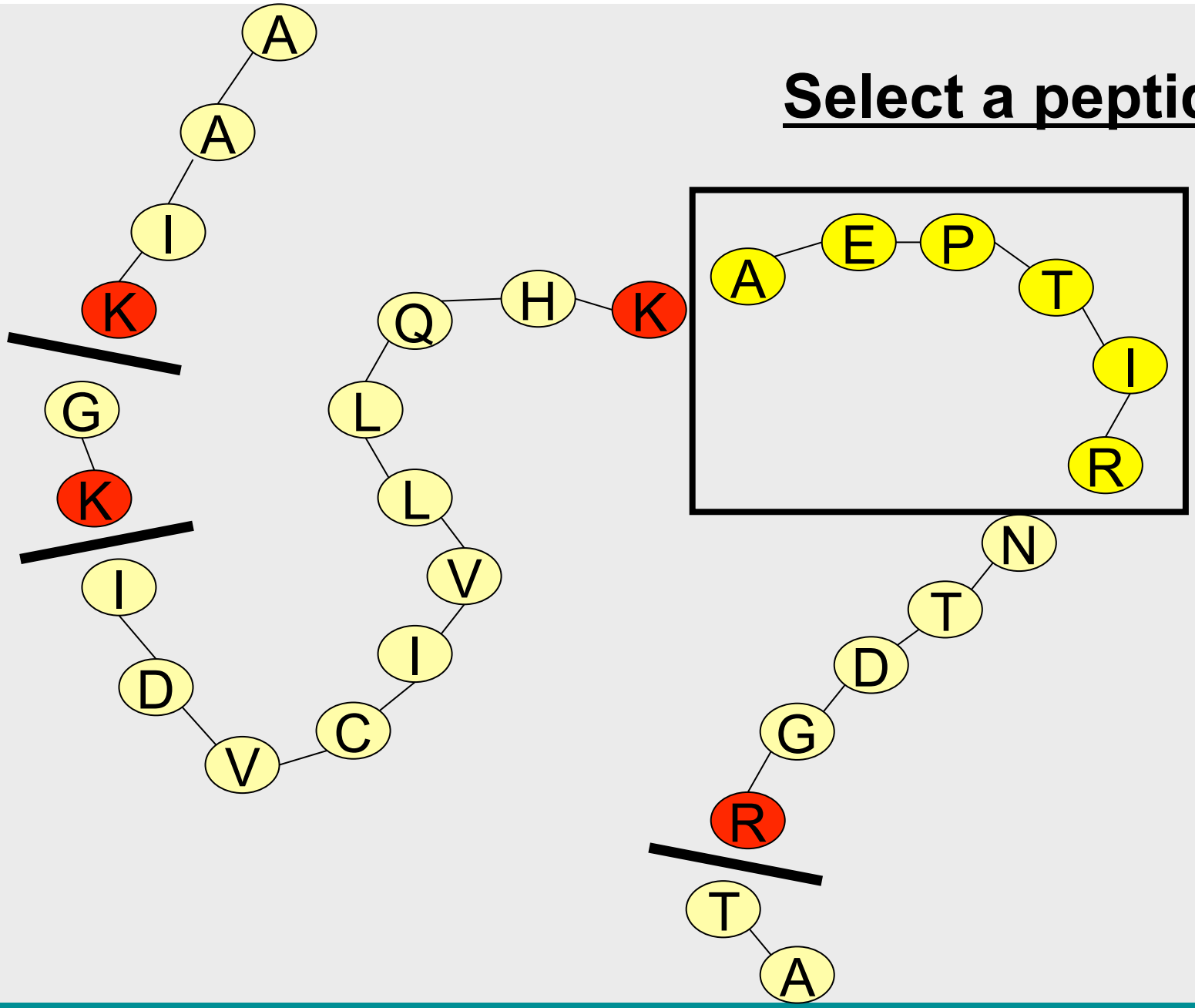
Start with a protein



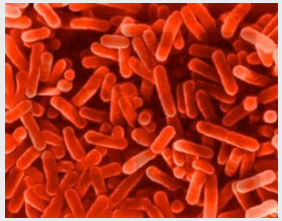
Cut with a protease (trypsin)



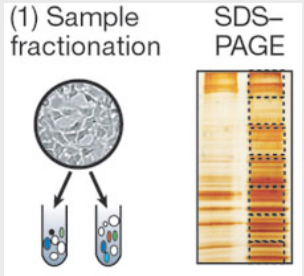
Select a peptide

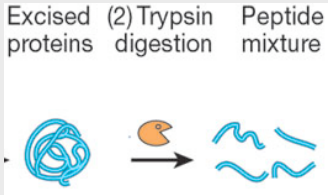


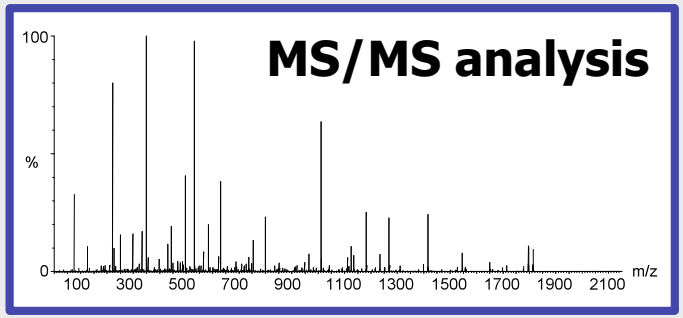
MS proteomics: overall workflow



PROTOCOL





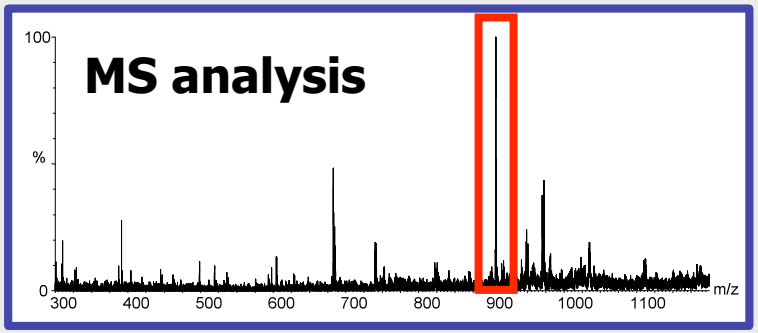


peptides

sequence database

proteins

fragmentation



THE RATIONALE BEHIND SHARING PROTEOMICS DATA

Juan A. Vizcaíno
juan@ebi.ac.uk

Need of data sharing in the proteomics field

EDITORIAL

nature biotechnology

Credit where credit is overdue

A universal tagging system that links data sets with the author(s) that generated them is essential to promote data sharing within the proteomics and other research communities.

Science progresses most rapidly when researchers provide access to their data. This is not only good scientific practice. It facilitates the confirmation of original results. It provides others with a starting point to explore new or related hypotheses. It speeds the identification of errors and discourages fraud. And it minimizes inefficient use of funding in duplicating experiments. And yet, full data disclosure in proteomics, and many other fields, remains a work in progress. If practicing scientists are to be truly incentivized to spend time and effort on sharing data, funders and publishers need to develop a universally recognized tagging system that would link investigators to their deposited data. In this way, publicly disclosed data sets would become part of a researcher's publication record, allowing such efforts to be recognized by employers and funders alike.

Next month marks the two-year anniversary of the publication of guidelines specifying the minimum reporting requirements for papers describing proteomics and molecular interaction experiments (*Nat. Biotechnol.* 25, 887–893, 894–898, 2007). Both sets of standards encourage deposition of data in public repositories, a practice that at the time was not universally adopted in proteomics.

We have carried out an informal survey of all manuscripts published in the year following publication of the two guidelines by the 68 authors of those two papers. The analysis reveals that a majority of the guideline authors published at least one manuscript last year for which no accompanying data were archived. If the proponents of data-reporting guidelines—most of whom are better resourced than other researchers in the field—are not depositing all of their data in a public repository, it is unlikely that the wider community is doing so either.

One issue that inhibits openness is the perception that full data disclosure may result in the loss of an edge over competing research groups. Occasionally, data are withheld while intellectual property is secured. More often, though, a failure to share simply reflects the considerable time and effort associated with formatting, documenting, annotating and releasing data. In this regard, the availability of new tools, such as an application (p. 598) to facilitate deposition of data in PRIDE (a public archive for mass spectrometry and protein identification data) should prove helpful.

For proteomics, the rapidly evolving technology and the complexity of the data itself pose particular challenges. Concerns about the quality of proteomics data generated by mass spectrometry have long plagued the field, raising the issue of whether peers have sufficient faith in other groups' work to not only value the data lodged in public repositories but also make the effort to deposit their own. Here too, though, progress is being made. A study reported in this issue (p. 633) demonstrates the high reproducibility of a targeted proteomic approach for biomarker discovery from plasma among several laboratories. Such a result would have been difficult to achieve using the technology and approaches of a few years ago.

But data quality is only part of the problem in overcoming the community's reticence about disclosure. For many researchers, the software provided by the public repositories for searching and analyzing proteomics data is not as efficient and user friendly as it could be. An analysis published last month by the Human Proteomics Organization cited the misassignment of peptides to ambiguously annotated proteins by database search engines as one of the major hindrances to researchers in the field (*Nat. Methods* 6, 423–430, 2009). What's more, despite the recent launch of yet another archive for mass spectrometry and protein identification data—the US National Center for Biotechnology Information's Peptidome repository (p. 600)—the various proteomics databases have yet to introduce a standardized data format that would allow the seamless exchange of data. Contrast this with the genome databanks, where the pooling of nucleotide sequence data in a common format has been integral to consistency, accessibility and, above all, utility of sequence data for reanalysis.

With all of these impediments, it's not surprising that proteomics researchers have been slow to embrace data disclosure. It is equally clear that disclosure edicts and recommendations from funding agencies and scientific journals have been insufficient to ensure widespread proteomics data release, despite evidence that the papers of researchers who share their data have an increased citation rate (*PLoS ONE* 2, e308, 2007). Clearly, other incentives are needed.

One option would be to provide researchers who release data to public repositories with a means of accreditation. This would take the form of a universally standardized tag for data that could be searched and recognized by both funding agencies and employers. An ability to search the literature for all online papers that used a particular data set would enable appropriate attribution for those who share. In essence, the tag would be a digital object identifier (DOI), currently best known for its use in unambiguously identifying papers online.

Similar to citation information about publications, citation information about a researcher's data DOIs could be gathered by funders assessing future support and used by institutions in performance evaluation. Researchers who disclose data sets that subsequently prove particularly useful to the community would end up with highly cited data DOIs, and could thereby be rewarded accordingly.

Such a system would not solve all the problems slowing data disclosure in proteomics and elsewhere. But it would provide greater incentive than the present system of evaluation, which is skewed almost exclusively to publications in high-profile journals and citation metrics. Data DOIs would not only enhance a researcher's reputation but also establish priority of data generation. Most important of all, they would provide a way to acknowledge the time and effort individuals must invest in sharing data, which ultimately benefits the scientific community as a whole.

Juan A. Vizcaíno
juan@ebi.ac.uk

Proteomics data sharing: why?

- 1) Data producers are not always the best data analysts
  Sharing of data allows analysts access to real data, and in turn allows better analysis tools to be developed

- 2) Meta-analysis of data can recycle previous findings for new tasks
  Putting findings in the context of other findings increases their scope

- 3) Sharing data allows independent review of the findings
  When actual replication of an experiment is often impossible, a re-analysis or spot checks on the obtained data become vitally important

- 4) Direct benefit for the field: fragmentation models, spectral libraries, ...

Simply sharing data is not enough…

Table 1. Identities of stress-induced protei

Spot ID	Synonym	Function
1202	SCO0525	Hypothetical p
3307	SCO2988	UDP-glucose
3509	SCO2180	Putative dihyc
6413	SCO6027	Probable acet
6419	SCO1494	3-Dehydroquit
6823	SCO5477	Putative oligo
118	SCO1340	Conserved hy
1104	SCO2368	Conserved hy
1617	SCO5373	ATP synthase
2601	SCO5373	ATP synthase
3616	SCO5371	ATP synthase
5721	SCO4814	Bifunctional p
1515	SCO2180	Putative dihyc
1616	SCO3661	Putative chap
2706	SCO3671	Heat shock pr
2906	SCO5999	Aconitase
3504	SCO1936	Putative trans
5310	SCO0506	NH(3)-depenc
7417	SCO5477	Putative oligo
505	SCO1998	30S ribosoma
1711	SCO1352	Xaa-pro amin
2618	SCO0681	Putative ferre
2722	SCO1998	30S ribosoma
4407	SCO5113	Oligopeptide
4509	SCO2390	Beta-ketoacyl
1803	SCO2181	2 Oxoglutarat
2113	SCO4277	Hypothetical p
3101	SCO3899	Hypothetical p
4309	SCO1081	Putative elect
4512	SCO5212	3-Phosphoshi
5514	SCO3629	Putative aden

+ SOD1

Table 1. *Identification of exosomal proteins based on MALDI-TOF peptide mass fingerprinting or MS/MS-derived sequences*

Band (Fig. 1)	Protein Name	Identification Method[a]	Accession Number[b]	Molecular Mass (kDa)	Matching Peptides	Sequence Coverage (%)
A	Mac-1 α-chain = CD11b		Théry et al., 1999 (14)			
1	Complement C3[c]	MS/MS (7)	4093220			
1	PK-120[c]	MS/MS (3)	Not in databases			
1	α2-Macroglobulin[c]	MS/MS (2)	Not in databases			
2	Plasminogen[c]	MS	P06868	91	28	37
3	Alix	MS MS/MS (6)	6755002[d]	96	26	34
3	Mac-1 β-chain = CD18	MS MS/MS (1)	P11835	85	27	38
4	hsp90-β = hsp84	MS	P11499	83	30	38
5	Serum albumin[c]	MS MS/MS (3)	P02769	69	42	66
B	hsc73		Théry et al., 1999 (14)			
B, C	MFG-E8/lactadherin		Théry et al., 1999 (14)			
6	Tubulin β	MS	P05218	50	20	44
7	Annexin VII = synexin	MS MS/MS (3)	Q07076	50	6	13
7	Bovine coagulation factor X[c]	MS/MS (2)	P00743	54		
7	PEDF[c]	MS/MS (3)	Q95121	46		
7	Tumor susceptibility protein (tsg) 101	MS/MS (2)	3184260[d]	44		
7	Rab GDP dissociation inhibitor (GDI) 3	MS MS/MS (1)	Q61598[e]	51	10	21
7	Elongation factor (EF) 1-α-1	MS/MS (2)	P10126	50		
7	EIF-4A-II	MS/MS (2)	P10630	46		
8	Annexin I	MS MS/MS (2)	P10107	39	7	25
8	Reverse transcriptase/pol (murine leukemia virus)	MS/MS (1)	61790[d]			
D	γ-Actin		Théry et al., 1999 (14)			
E	G protein $G_{i2}\alpha$ subunit		Théry et al., 1999 (14)			
F	Annexin II		Théry et al., 1999 (14)			
9	Annexin V	MS MS/MS (4)	P48036	36	16	54
10	Annexin IV	MS MS/MS (4)	P97429	36	20	63
10	Galectin-3 = Mac-2	MS MS/MS (6)	P16110	27	11	37
11	Syntenin	MS MS/MS (6)	2197106[d]	32	17	35
G	Gag polyprotein (murine leukemia virus)		Théry et al., 1999 (14)			
G	MHC class II β-chain		Théry et al., 1999 (14)			
12	14-3-3 protein η	MS MS/MS (4)	P11576	28	21	68
12	14-3-3 protein γ/δ	MS MS/MS (2)	P35215	28	20	63
12	14-3-3 protein γ	MS/MS (1)	3065929[d]			
13	Apolipoprotein A-I[c]	MS	P15497	30	25	67
H	CD9		Théry et al., 1999 (14)			
14	Thioredoxin peroxidase II	MS MS/MS (6)	P35700	22	8	43
14	Rab 11	MS/MS (1)	P46638	24		
14	κ-Casein[c]	MS/MS (2)	P02668	21		
15	Rab-7	MS MS/MS (3)	P51150	24	5	26
16	Ferritin light chain[c]	MS	O46415	20	15	73
16	Rap1-B	MS	P09526	21	14	57
17	Cofilin	MS	P18760	19	10	50
18	Histone H3	MS	Z85979[e]	15	7	45
19	Histone H2B	MS	P10853	14	13	82
19	Histone H2A	MS	P20670	14	12	67
20	Histone H4	MS	90626[d]	11	15	90
20	Profilin I	MS	P10924	15	11	60
21	Hemoglobin γ-chain[c]	MS	P02081	16	16	74
21	Hemoglobin α-chain[c]	MS	P01966	15	9	66

Stress	(kDa)/pI Theoretical	Accession no.	Species
CS	69.07/5.8	BAA97338	*Arabidopsis thaliana*
CS			
CS	62.82/6.4	P42863	*Oryza sativa*
CS			
CS	65.29/6	Q43097	*Lotus japonicus*
CS			
EtOH	62.82/6.4	P42863	*Oryza sativa*
EtOH			
EtOH	67.71/6.9	Q00775	*Salanum tuberosum*
EtOH			
EtOH	67.69/7.1	BAA77351	*Triticum aestivum*
EtOH			
HS	62.64/8.5	Q42608	*Brassica rapa*
HS			
HS	54.96/5.2	Q38681	*Acetabularia acetabulum*
HS			
HS	57.44/7	P30567	*Gossypium hirsutum*
HS			
HS	57.94/8	P37215	*Lycopersicon esculentum*
NaCl			
NaCl			
NaCl	49.59/7.1	S33520	Soybean
NaCl	43.04/6.1	P51110	*Lycopersicon esculentum*
NaCl			
NaCl	38.79/6.2	P51110	*Vitis vinifera*
P1	27.54/8.8	BAB03428	*Oryza sativa*
P1	27.54/8.8	BAB03428	*Oryza sativa*
P1	24.36/8.6	BAA92870	*Oryza sativa*
P1	26.58/6.4	P09886	*Pisum sativum*
P1			
P1			
P1	56.77/6.1	P55238	*Hordeum vulgare*

A nuance: *available* data vs. *accessible* data

When data is only made available as arbitrarily formatted tables, it carries important limitations

- Source data are not made available
  - No peer review validation possible
  - Very little raw materials for testing innovative *in silico* techniques are available
  - Traceability of data is lost quickly in downstream results
- Automated (re-)processing of the results (e.g., identifications) is impossible
- Data producers do not actually feed their results and knowledge back to the community

Community standards for proteomics

The Human Proteome Organisation (HUPO)
Proteomics Standards Initiative (PSI)

http://www.psidev.info

- **Creates minimal requirements, standard formats, and CV's and ontologies**
- **Composed of several workgroups**

Molecular Interactions	***(MI)***	***PSI-MI format v2.5***
Mass Spectrometry	***(MS)***	***mzData, mzML format***
Protein Separation	***(PS)***	***GelML format***
Proteomics Informatics	***(PI)***	***mzIdentML format***
Protein Modifications	***(Mod)***	***PSI-MOD ontology***

How do we make this all happen?

- Journal guidelines

  Journal guidelines heavily influence the decisions taken by authors; by first requesting and subsequently mandating data submission to established repositories, they provide an important stick.

- Funder support and guidelines

  Funders contribute both sticks and carrots. The sticks lie in the grant application guidelines; they can require a plan for data management and dissemination. The carrot is in providing specific funding for this aspect of science.

- Data repositories

  The availability of reliable, freely available repositories is key; submission thresholds should be kept low and **added value** needs to be provided. Furthermore, feedback loops need to be established in order to ensure that accumulated data flows back to the user community. Repositories thus provide mostly carrots.

PROTEOMICS DATA REPOSITORIES AVAILABLE TODAY

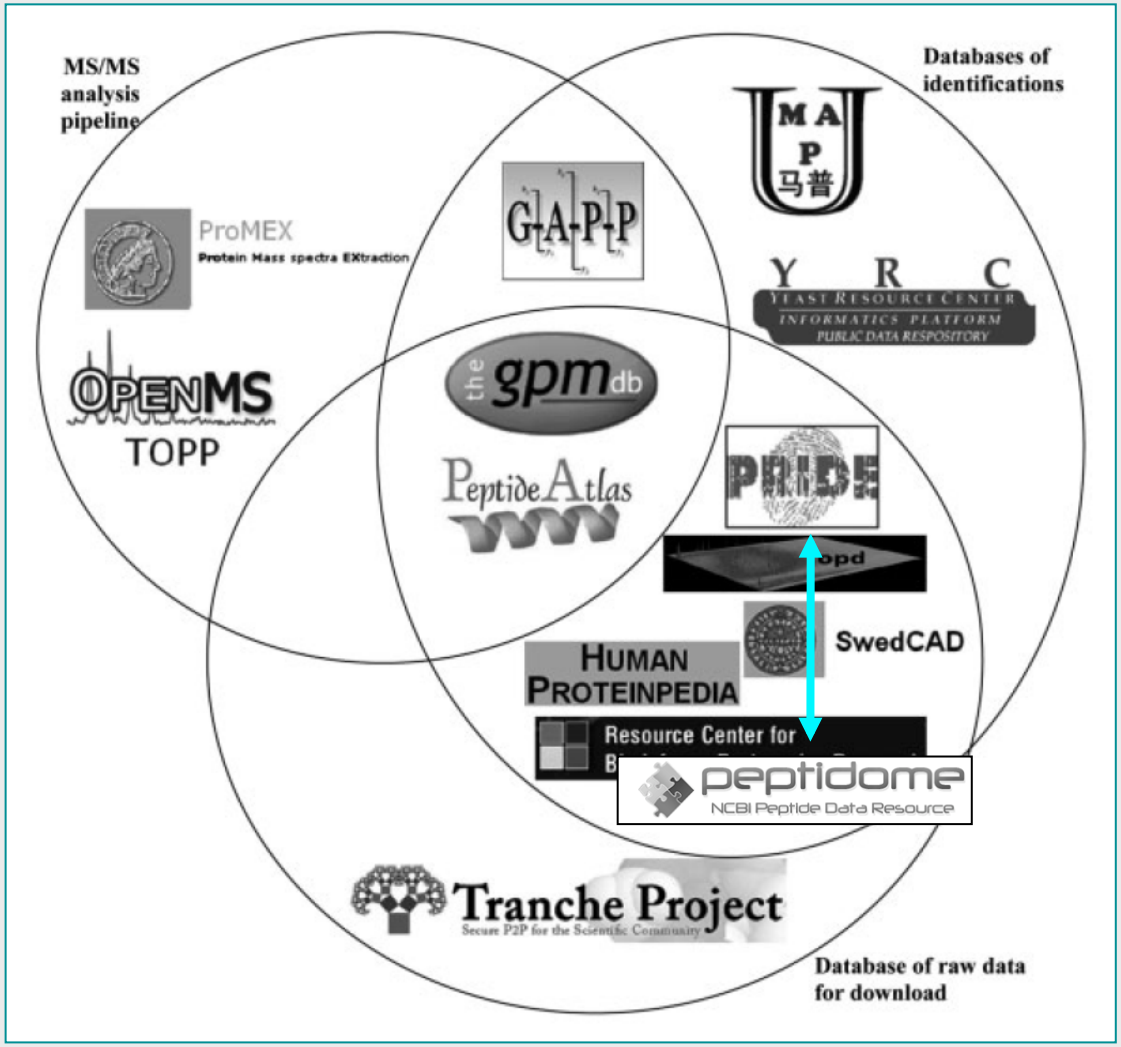
Existing proteomics repositories

- Main public repositories:
  - PROteomics IDEntifications database (PRIDE)
  - Global Proteome Machine (GPMDB)
  - Peptide Atlas
  - Tranche
  - NCBI Peptidome

- Smaller scale repositories, more specialized:
  Among others: Human Proteinpedia, Genome Annotation Proteomics Pipeline (GAPP), MAPU, SwedCAD, PepSeeker, Open Proteomics Database, …

- Very diverse: different aims, functionalities, …

A comprehensive view on existing systems



From: Mead *et al.*, *Proteomics*, 2009

Types of information stored

- 1) **Original experimental data** recorded by the mass spectrometer (primary data)

Primary data

Binary data

XML-based files

mzData

mzXML

mzML

Peak lists

.dta, .pkl, .mgf, .ms2

Types of information stored

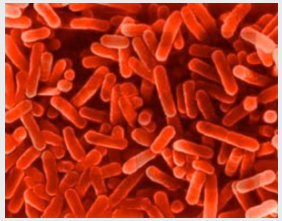
- 1) **Original experimental data** recorded by the mass spectrometer (primary data)
- 2) **Identification results** inferred from the original primary data

Peptide and Protein Identifications

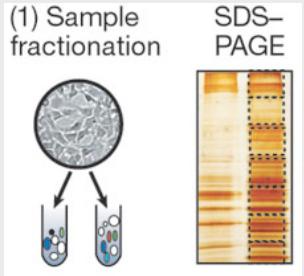
mzIdentML,
mascot .dat,
sequest .out,
SpectrumMill .spo
pep.xml, prot.xml

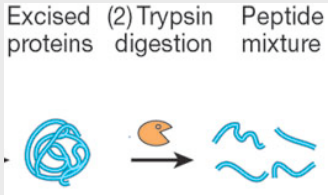
Only qualitative data!

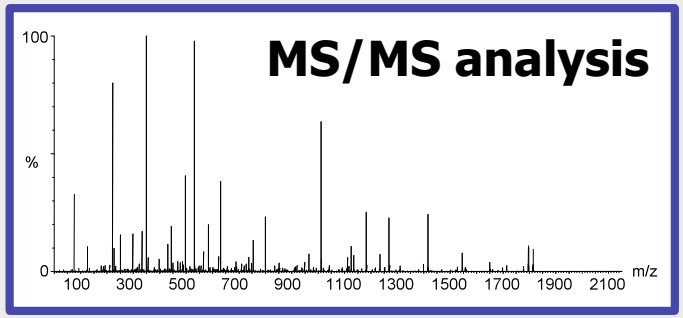
MS proteomics: overall workflow



PROTOCOL





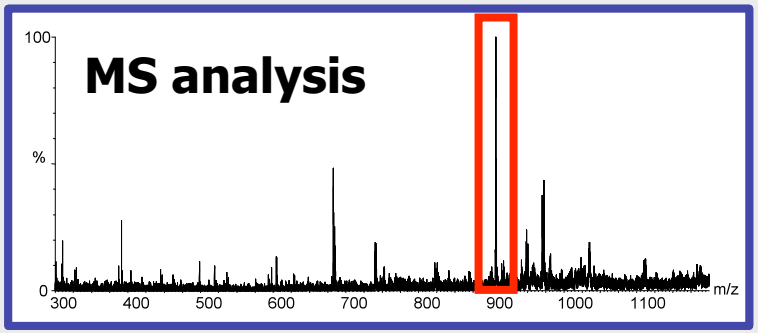


peptides

sequence database

proteins

fragmentation



Intermezzo: Protein inference

The *minimal* and *maximal* explanatory sets

peptide	*a*	*b*	*c*	*d*
proteins				
prot X	*x*		*x*	
~~*prot Y*~~	~~*x*~~			
prot Z		*x*	*x*	*x*

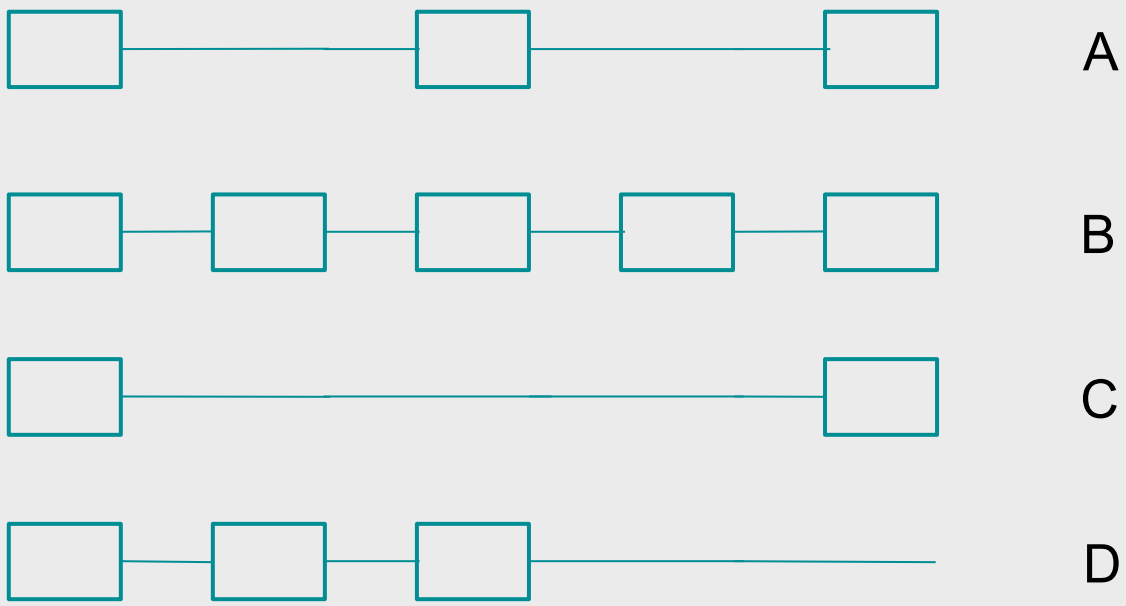
Minimal set
Occam

The Truth

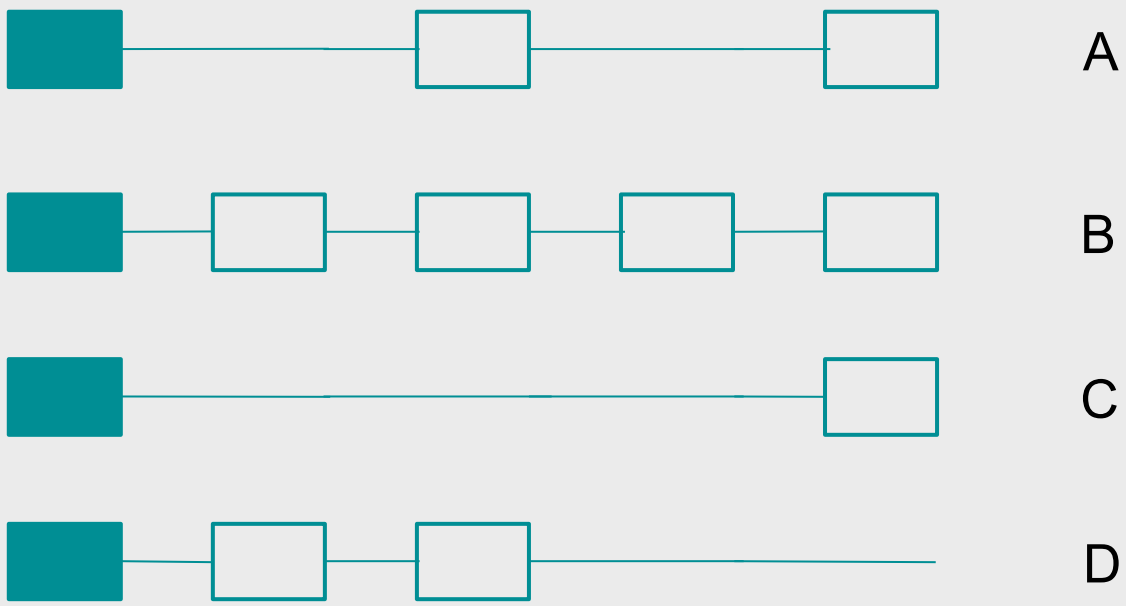
peptide	*a*	*b*	*c*	*d*
proteins				
prot X	*x*		*x*	
prot Y	*x*			
prot Z		*x*	*x*	*x*

Maximal set
anti-Occam

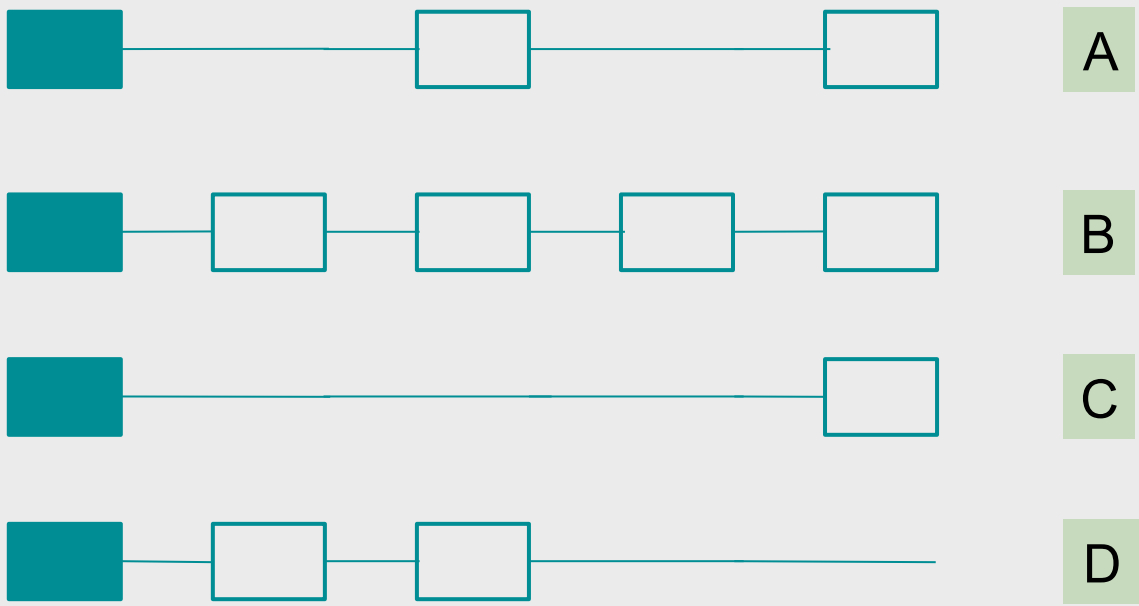
An additional layer of complexity…



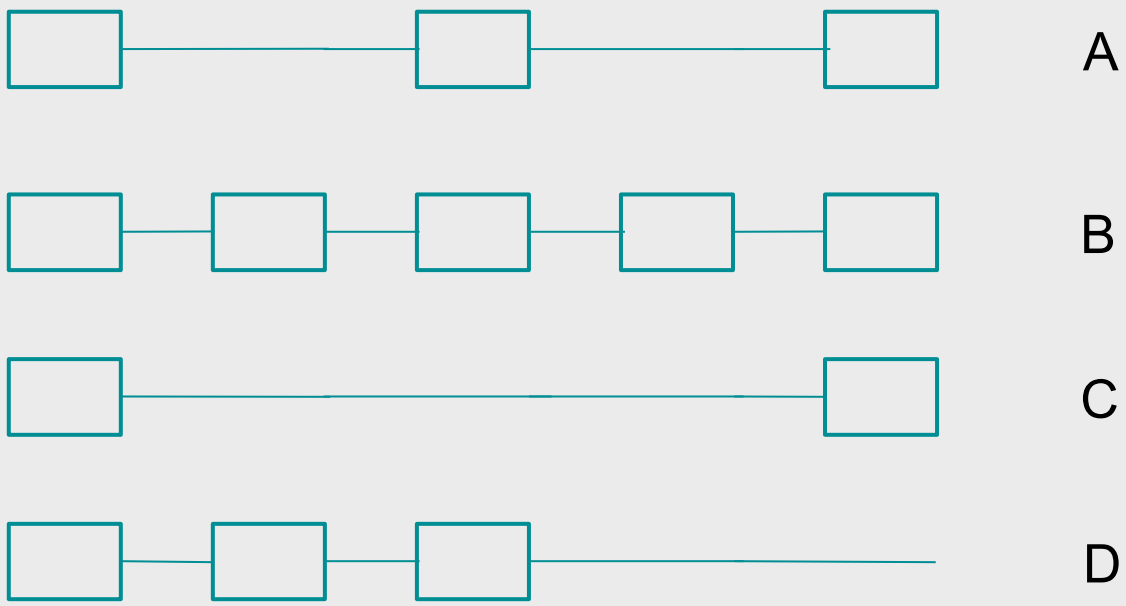
Protein inference



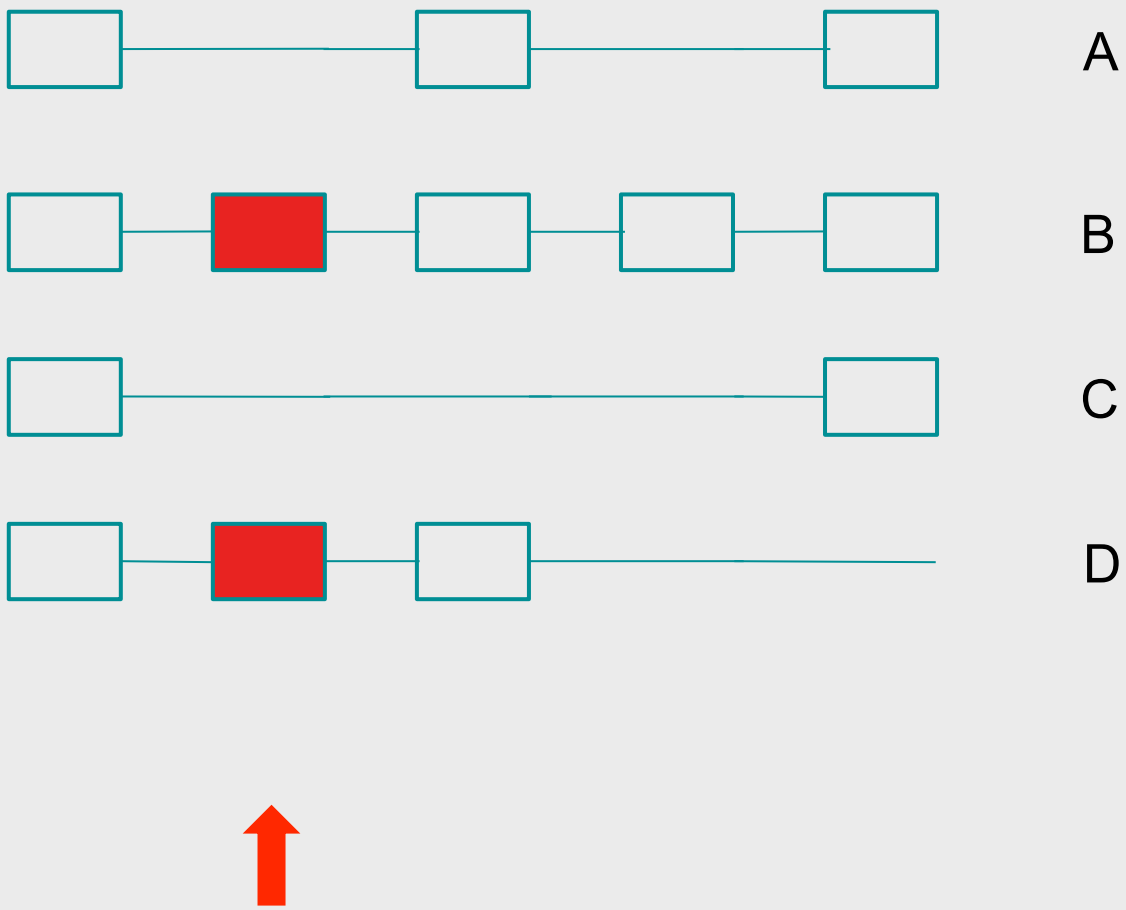
Protein inference



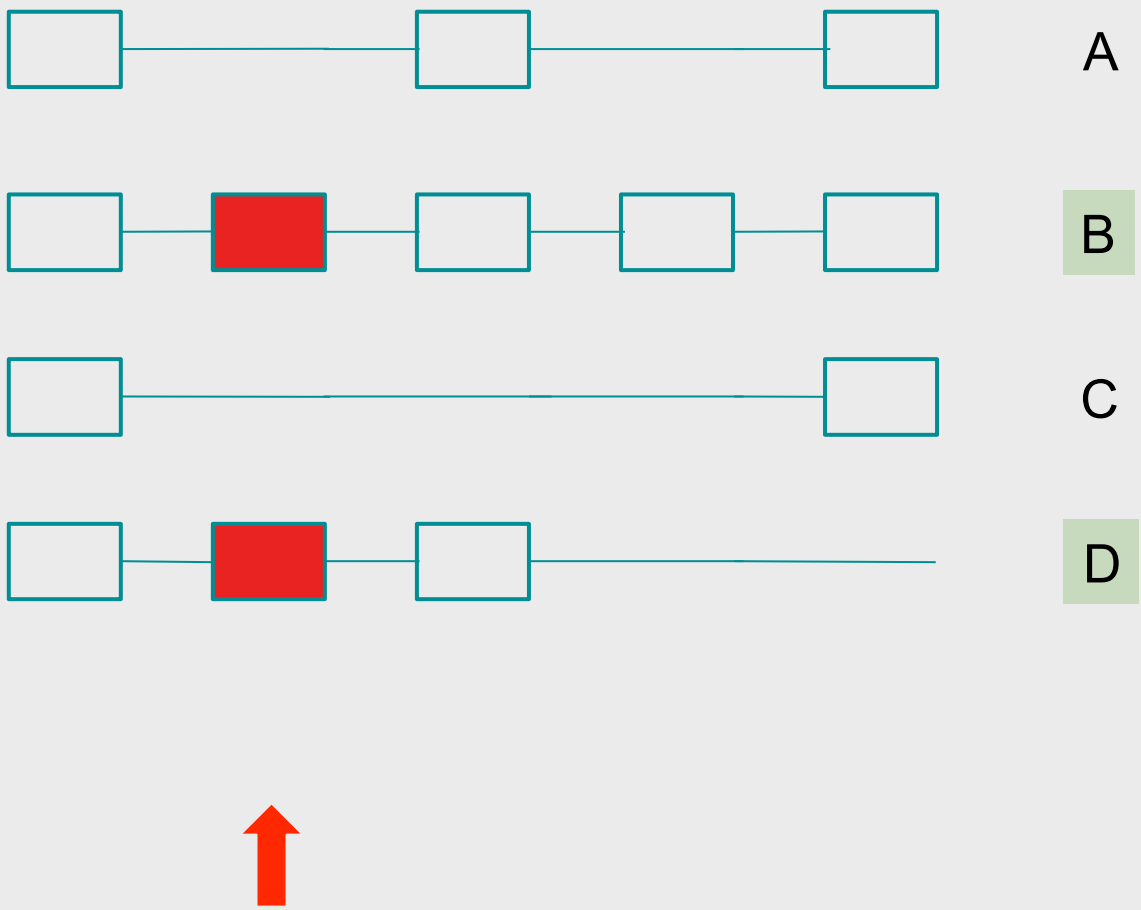
Protein inference



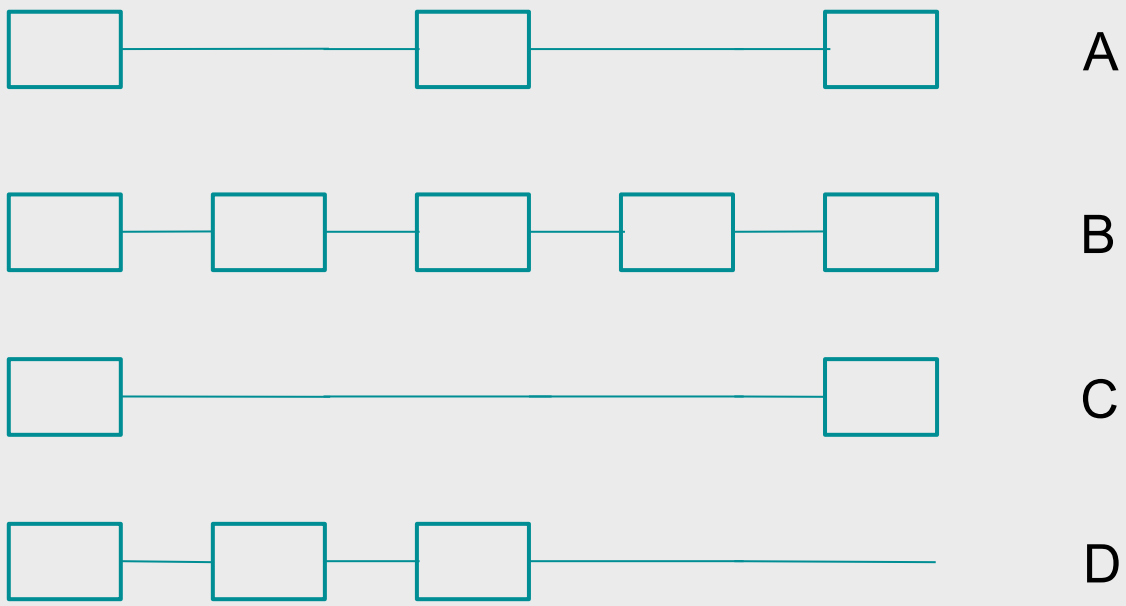
Protein inference



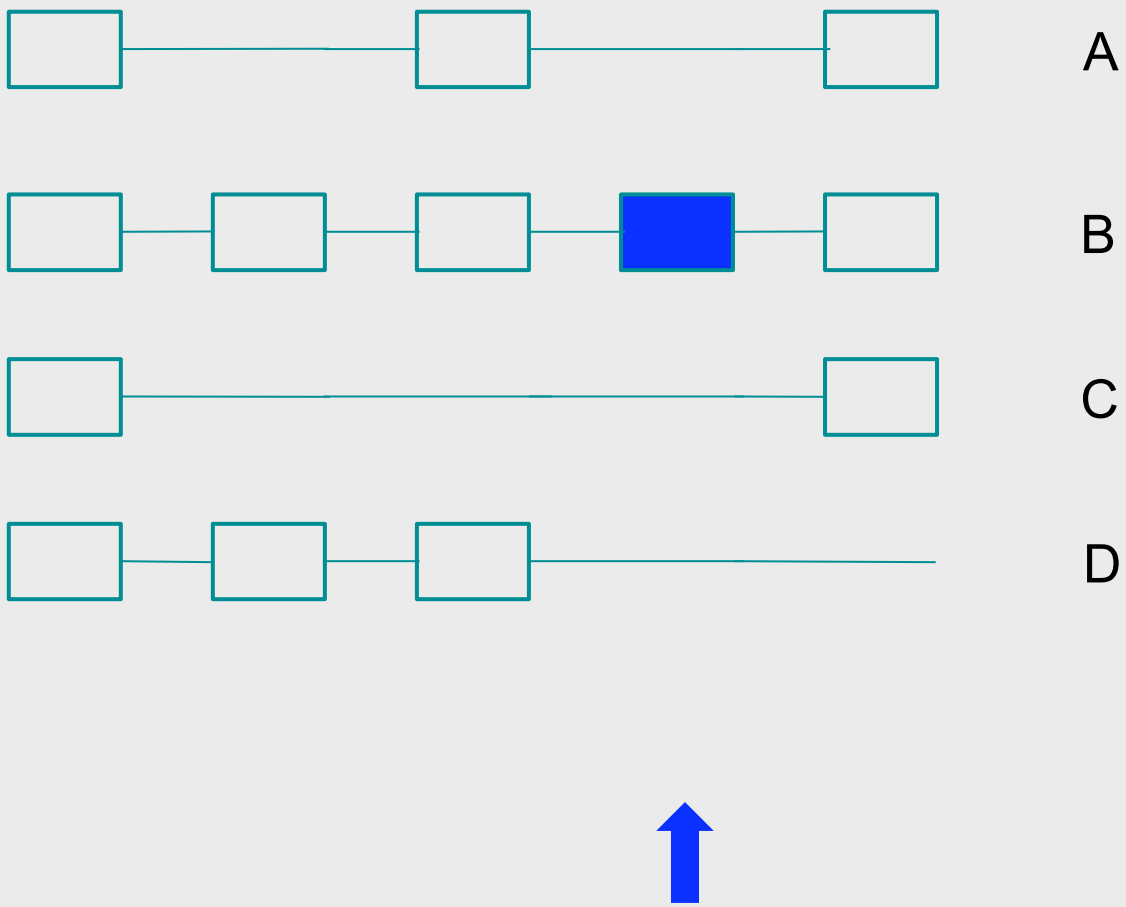
Protein inference



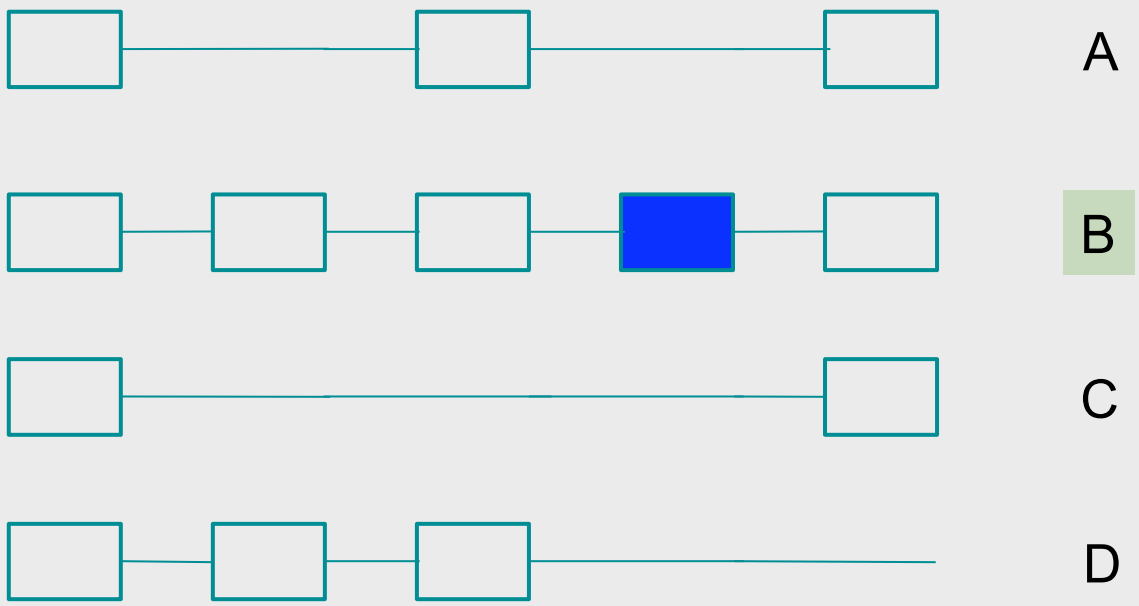
Protein inference



Protein inference



Protein inference



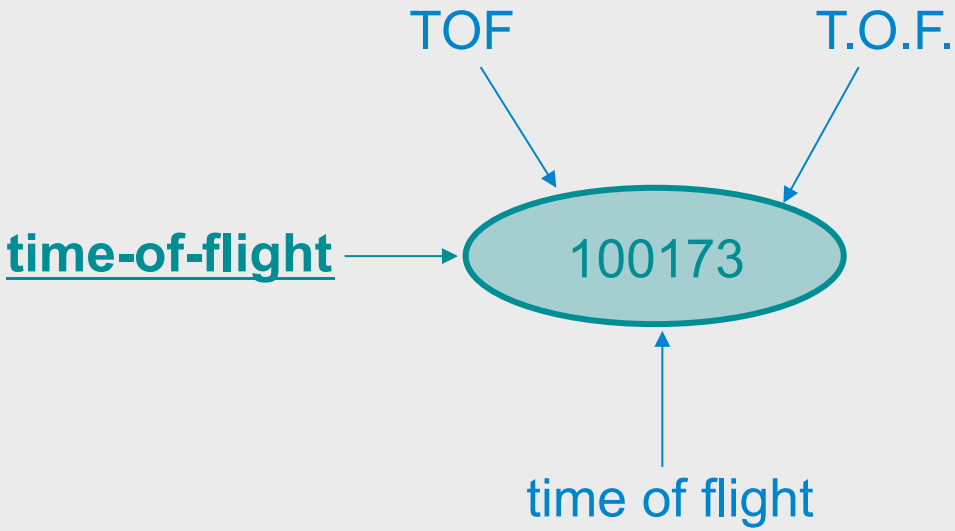
Unambiguous peptide

Types of information stored

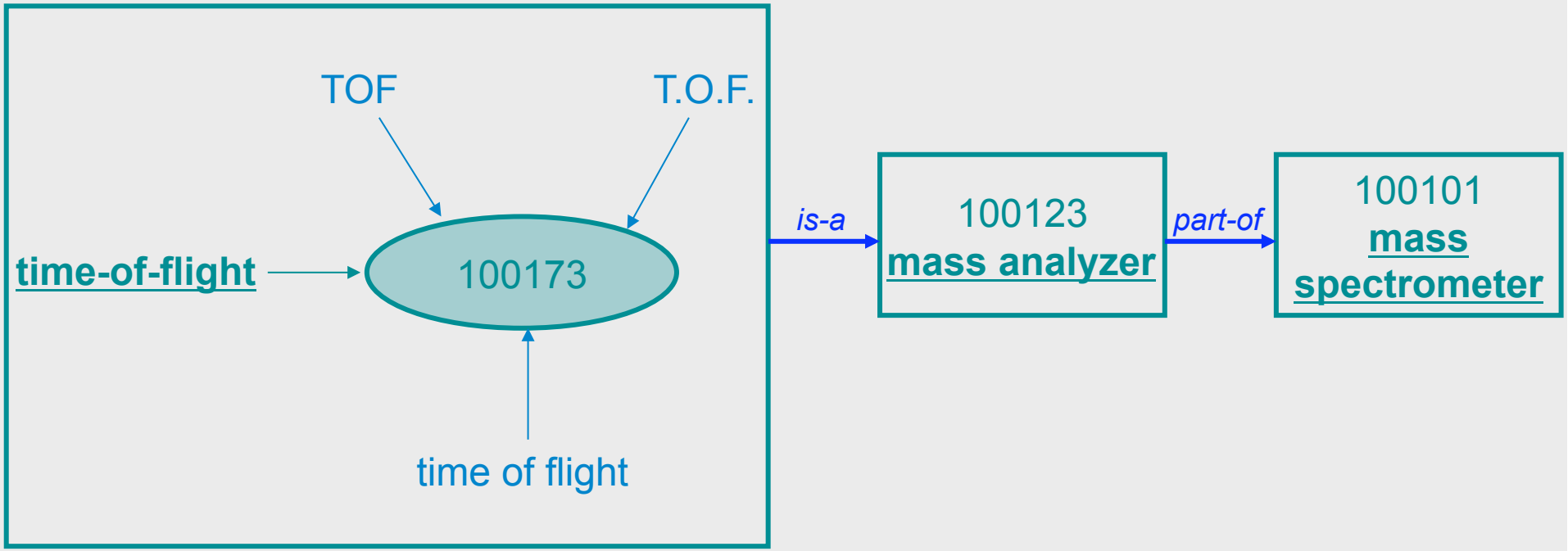
- 1) **Original experimental data** recorded by the mass spectrometer (primary data)
- 2) **Identification results** inferred from the original primary data
- 3) Experimental and technical **metadata**

Controlled Vocabularies (CVs)

Term
Synonyms



Relationships between CV terms



CVs, ontologies (here: PSI-MOD)

http://www.ebi.ac.uk/ols

OLS - Ontology Lookup Service

MOD Ontology Browser

- protein modification
  - uncategorized protein modification
  - protein modification categorized by isobaric sets
  - protein modification categorized by chemical process
    - alkylated residue
    - crosslinked residues
    - deamidated residue
    - cyclized residue
    - acylated residue
      - acetylated residue
        - N-acetylated residue
          - N-acetyl-L-alanine
          - N-acetyl-L-aspartic acid
          - N-acetyl-L-cysteine
          - N-acetyl-L-glutamic acid
          - N-acetyl-L-glutamine
          - N-acetylglycine
          - N-acetyl-L-isoleucine
          - N-acetyl-L-methionine
          - N-acetyl-L-proline
          - N-acetyl-L-serine
          - N-acetyl-L-threonine
          - N-acetyl-L-tyrosine
          - N-acetyl-L-valine
          - N2-acetyl-L-arginine
          - N,O-diacetylated L-serine
          - N-acetylated L-lysine
            - N2-acetyl-L-lysine

Legend:
is a
develops from
part of
other

Help (hide)

Double-click a term to see its children. The ontology browser is populated dynamically. If the
children for a given term, there may be a small delay while the browser fetches **Click** to high
see any information associated with it. **Hover** over a term to see its relation with its immediat
terms will not display any relational information.

Relations

N-acetyl-L-serine is_a N-acetylated residue

Term Information

ID: MOD:00723

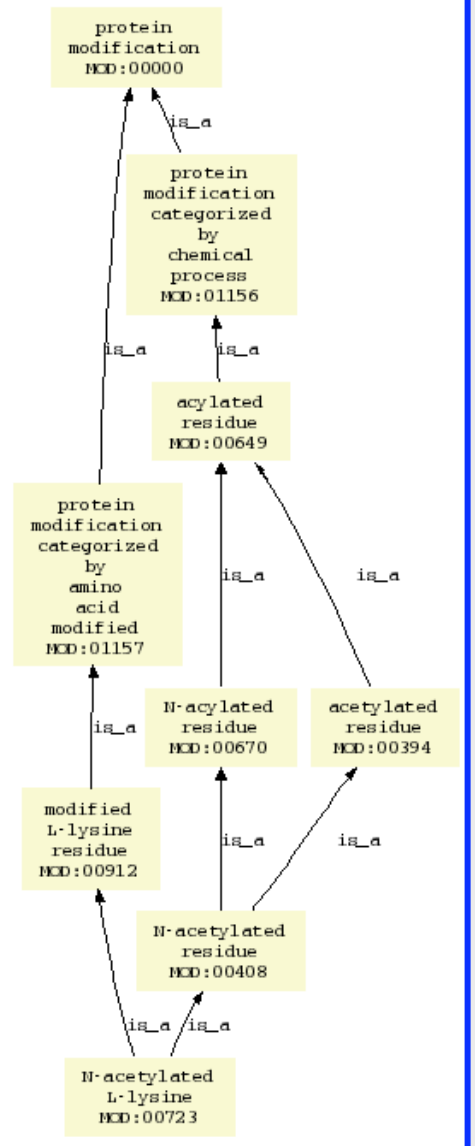
Name: N-acetylated L-lysine

Associated information

definition	A protein modification that effectively converts an L-lysine residue to eithe L-lysine, or N6-acetyl-L-lysine.
DiffAvg	42.04
DiffFormula	C 2 H 2 N 0 O 1
DiffMono	42.010565
Formula	C 8 H 15 N 2 O 2
MassAvg	171.22
MassMono	171.113353
Origin	K
Source	Natural
TermSpec	none
preferred name	N-acetylated L-lysine
exact synonym	AcLys
xref_definition	PSI-MOD:ref

Term Hierarchy

Paths to Root: Child relationships:



Types of information stored

- 1) **Original experimental data** recorded by the mass spectrometer (primary data)
- 2) **Identification results** inferred from the original primary data
- 3) Experimental and technical **metadata**
- 4) **Quantitation** information

Wide variety of quantitative techniques…

Quantitation: Overview

Many different approaches to protein quantitation using mass spectrometry data have been described in the literature. For a short, recent review, see Ong, S. E. and Mann, M., Mass spectrometry-based proteomics turns quantitative, Nature Chemical Biology 1 252-262 (2005). In terms of the "mechanics" of their implementation, most of the popular approaches can be classified into a relatively small number of ***protocols***:

- **Reporter:** Quantitation based on the relative intensities of fragment peaks at fixed m/z values within an MS/MS spectrum. For example, iTRAQ and Tandem Mass Tags
- **Precursor:** Quantitation based on the relative intensities of extracted ion chromatograms (XICs) for precursors within a single data set. This is by far the most widely used approach, which can be used with any chemistry that creates a precursor mass shift. For example, ^{18}O, AQUA, ICAT, ICPL, Metabolic, SILAC, etc., etc.
- **Multiplex:** Quantitation based on the relative intensities of sequence ion fragment peaks within an MS/MS spectrum. This is a novel approach, which can be used with labels located at the peptide terminus, such as ^{18}O or SILAC at K or R in combination with tryptic digestion.
- **Replicate:** Label free quantitation based on the relative intensities of extracted ion chromatograms (XICs) for precursors in multiple data sets aligned using mass and elution time.
- **emPAI:** Label free quantitation for the proteins in a mixture based on protein coverage by the peptide matches in a database search result.
- **Average:** Label free quantitation for the proteins in a mixture based on the application of a rule to the intensities of extracted ion chromatograms (XICs) for the peptide matches in a database search result.

Some protocols can be fully implemented within a Mascot result report because all the necessary information is present in the peak list. These protocols are Reporter, Multiplex, and emPAI. In fact, emPAI is "always on", and will be reported whenever an MS/MS search contains at least 100 spectra.

The other three protocols require additional information from the raw data file, either because it is necessary to integrate the elution profile of each precursor peptide or because information is required for precursor peptides that were not used to trigger MS/MS scans, so are missing from the peak list. So, for Precursor, Replicate, and Average, the quantitation report is generated in Mascot Distiller, which has access to both the Mascot search results and the raw data.

Quantitation techniques

Label free

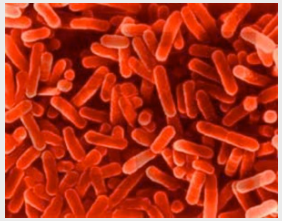
Gel-based quantitation approaches

-Different philosophies
-Very heterogeneous data formats
-Techniques not very well established

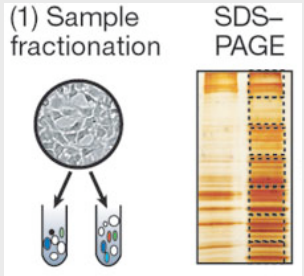
Very problematic data for proteomics repositories

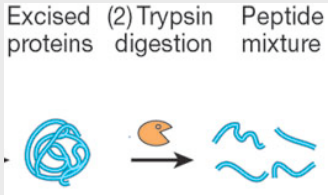
PRIDE

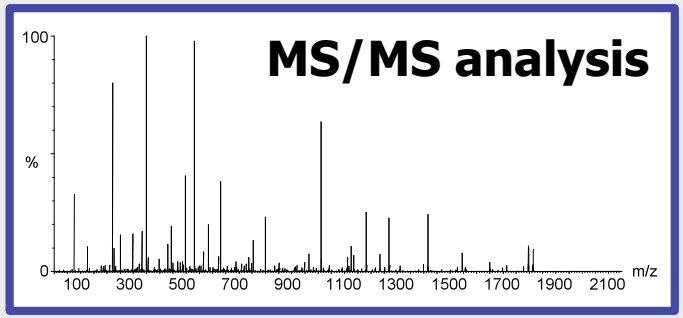
MS proteomics: overall workflow



PROTOCOL





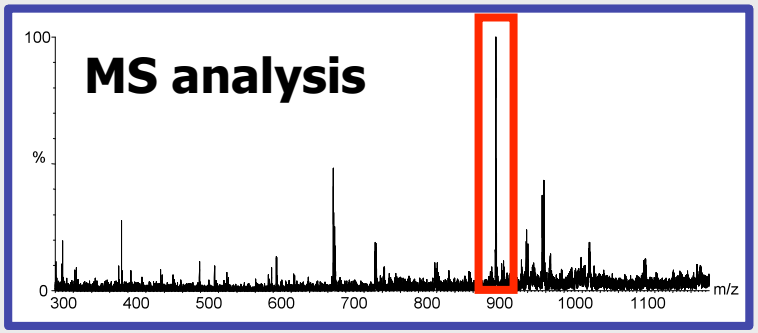


peptides

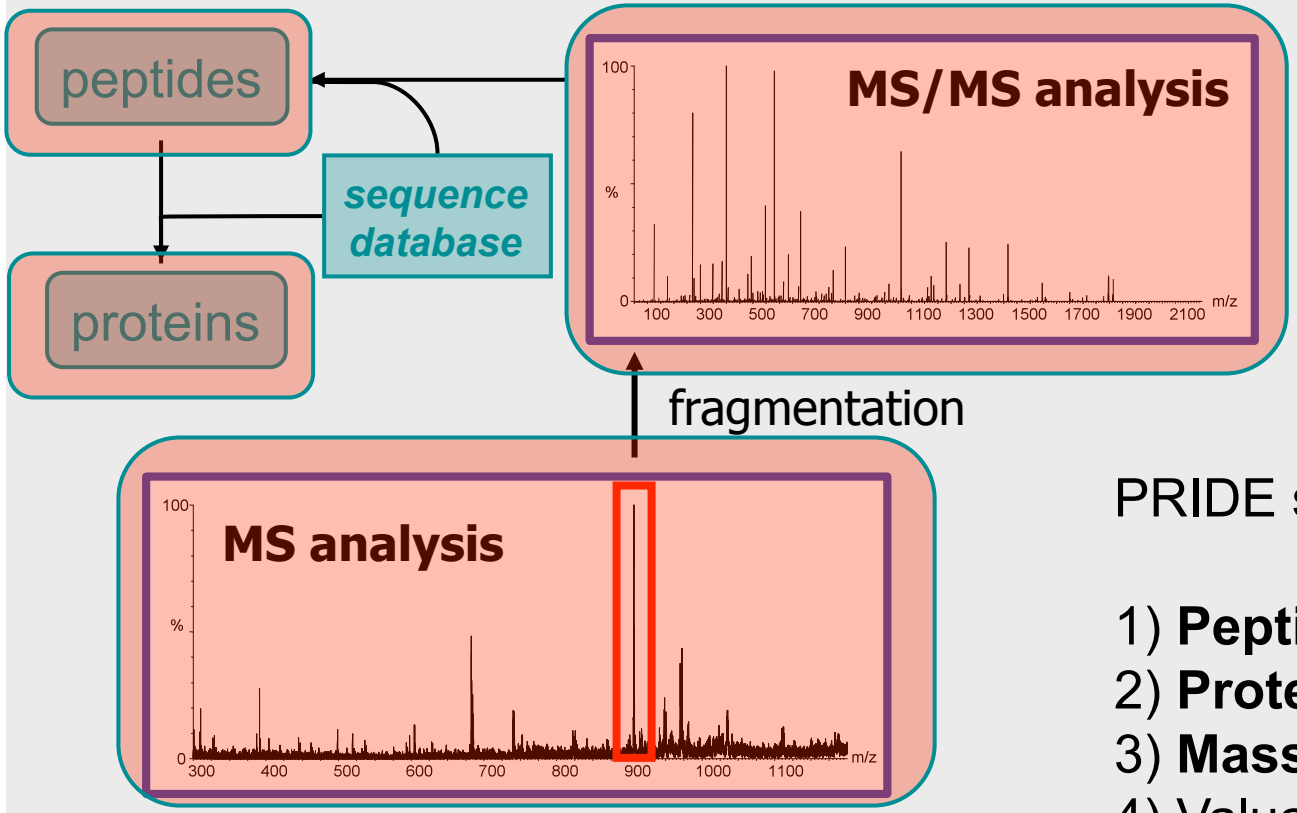
sequence database

proteins

fragmentation



PRIDE database (www.ebi.ac.uk/pride)



PRIDE stores:

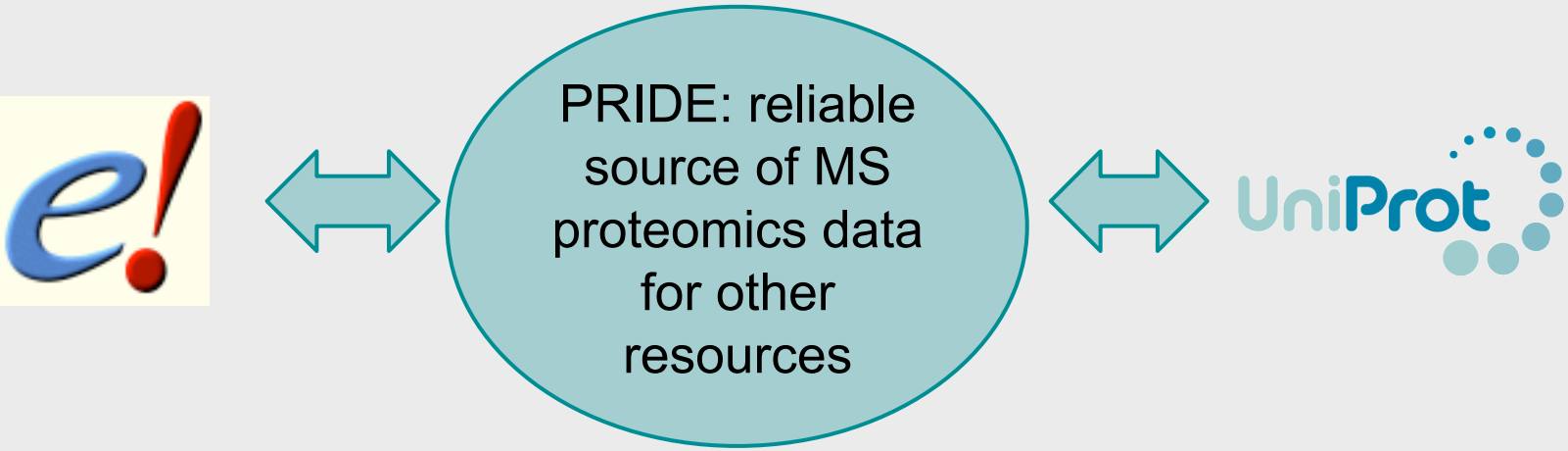
1) **Peptide IDs**
2) **Protein IDs**
3) **Mass spectra** as peak lists
4) Valuable additional **metadata**

PRIDE: why is it there?

- Repository to support publications (proteomics MS derived data)

PRIDE: why is it there?

- Repository to support publications (proteomics MS derived data)
- Source of proteomics data for other data resources



THE LOOK OF PRIDE

PRIDE web interface – overview

PRoteomics IDEntifications database (PRIDE)

PRIDE Basic Statistics

Search filtered on: Sample Parameters (Species, tissue, disease etc.) , with parameters Type: 'BTO', Value: 'BTO:0000142'

View Instructions

This Table Describes 88 Experiments.

To sort by any of the first seven columns, click the heading.
(Repeated clicking changes the direction of the sort.)

Compare Experiments

Accession	Title	Species	Tissue	Cell Type	GO Term	Disease	Protein Count	Peptide Count	Spectra Count	Retrieve Details (Output format set above.)	Compare Protein Identification Sets	Select Reference Experiment
1636	Proteomics Mapping of Brain Plasma Membrane Proteins	Mus musculus (Mouse)	cerebral cortex, hippocampus	-	-	-	2356	9510	0	Download	☐	
1637	Characterization of the Mouse Brain Proteome Using Global Proteomic Analysis Complemented with Cysteinyl-Peptide Enrichment (Protein ID Set 1)	Mus musculus (Mouse)	brain	-	-	DiseaseFree	599	21218	0	Download	☐	
1638	Characterization of the Mouse Brain Proteome Using Global Proteomic Analysis Complemented with Cysteinyl-Peptide Enrichment (Protein ID Set 2)	Mus musculus (Mouse)	brain	-	-	DiseaseFree	800	11078	0	Download	☐	
1639	Characterization of the Mouse Brain Proteome Using Global Proteomic Analysis Complemented with Cysteinyl-Peptide Enrichment (Protein ID Set 3)	Mus musculus (Mouse)	brain	-	-	DiseaseFree	1000	8255	0	Download	☐	
1640	Characterization of the Mouse Brain Proteome Using Global Proteomic Analysis Complemented with Cysteinyl-Peptide Enrichment (Protein ID Set 4)	Mus musculus (Mouse)	brain	-	-	DiseaseFree	1600	7470	0	Download	☐	

PART_OF

10116	Rattus norvegicus (Rat)	BTO:0000383	kidney tumor cell line
4932	Saccharomyces cerevisiae (Baker's yeast)	BTO:0000713	leaf
		BTO:0001629	left ventricle
602	Salmonella typhimurium	BTO:0000759	liver

PRIDE web interface – experiment and protein

Experiment View

Human CSF analysis (L

Accession:	1755
Short Label:	LCQ C
Additional Information:	Sour PRID

*SkyPainter is a tool to deter mapped UniProt accessions

Instrument:

Identification Detail View

Details for identification: IPI00400826.1

Submitted Accession	IPI00400826.1
Search Database	ipi.HUMAN
Cross-References	Accession / Database: ENSP00000315130 ENSEMBL_HUMAN; ENSP00000369812 ENSEMBL_HUMAN; IPI00400826.1 IPI; NP_001822.2 REFSEQ. These mappings have been obtained using the Protein Identifier Cross Reference (PICR) Service at the EBI. They are based on 100% sequence identity and, as a further requirement where applicable, all submitted peptide sequences must match. Mappings shown in light grey are historical and correspond to inactive entries in the source databases.
Search Engine	proteinprophet
Score	1.0
Treshold	0.9
Sequence Coverage	0.343

Additional

Source	Name	Value
PRIDE	Protein description line	Clusterin isoform 1
PRIDE	ProteinProphet probability score	1.0
PRIDE	Indistinguishable alternative protein accession	IPI00291262

Mapped Protein sequence

```
0001  MQVCSQPQRG CVREQSAINT APPSAHNAAS PGGARGHRVP LTEACKDSRI GGMMKTLLLF VGLLLTWESG QVLGDQTVSD  0080
0081  NELQEMSNQG SKYVNKEIQN AVNGVKQIKT LIEKTNEERK TLLSNLEEAK KKKEDALNET RESETKLKEL PGVCNETMMA  0160
0161  LWEECKPCLK QTCMKFYARV CRSGSGLVGR QLEEFLNQSS PFYFWMNGDR IDSLLENDRQ QTHMLDVMQD HFSRASSIID  0240
0241  ELFQDRFFTR EPQDTYHYLP FSLPHRRPHF FFPKSRIVRS LMPFSPYEPL NFHAMFQPFL EMIHEAQQAM DIHFHSPAFQ  0320
0321  HPPTEFIREG DDDRTVCREI RHNSTGCLRM KDQCDKCREI LSVDCSTNNP SQAKLRRELD ESLQVAERLT RKYNELLKSY  0400
0401  QWKMLNTSSL LEQLNEQFNW VSRLANLTQG EDQYYLRVTT VASHTSDSDV PSGVTEVVVK LFDSDPITVT VPVEVSRKNP  0480
0481  KFMETVAEKA LQEYRKKHRE E
```

Submitted Protein Sequence

The protein sequence that was submitted with this identification is identical to the current mapped protein sequence.

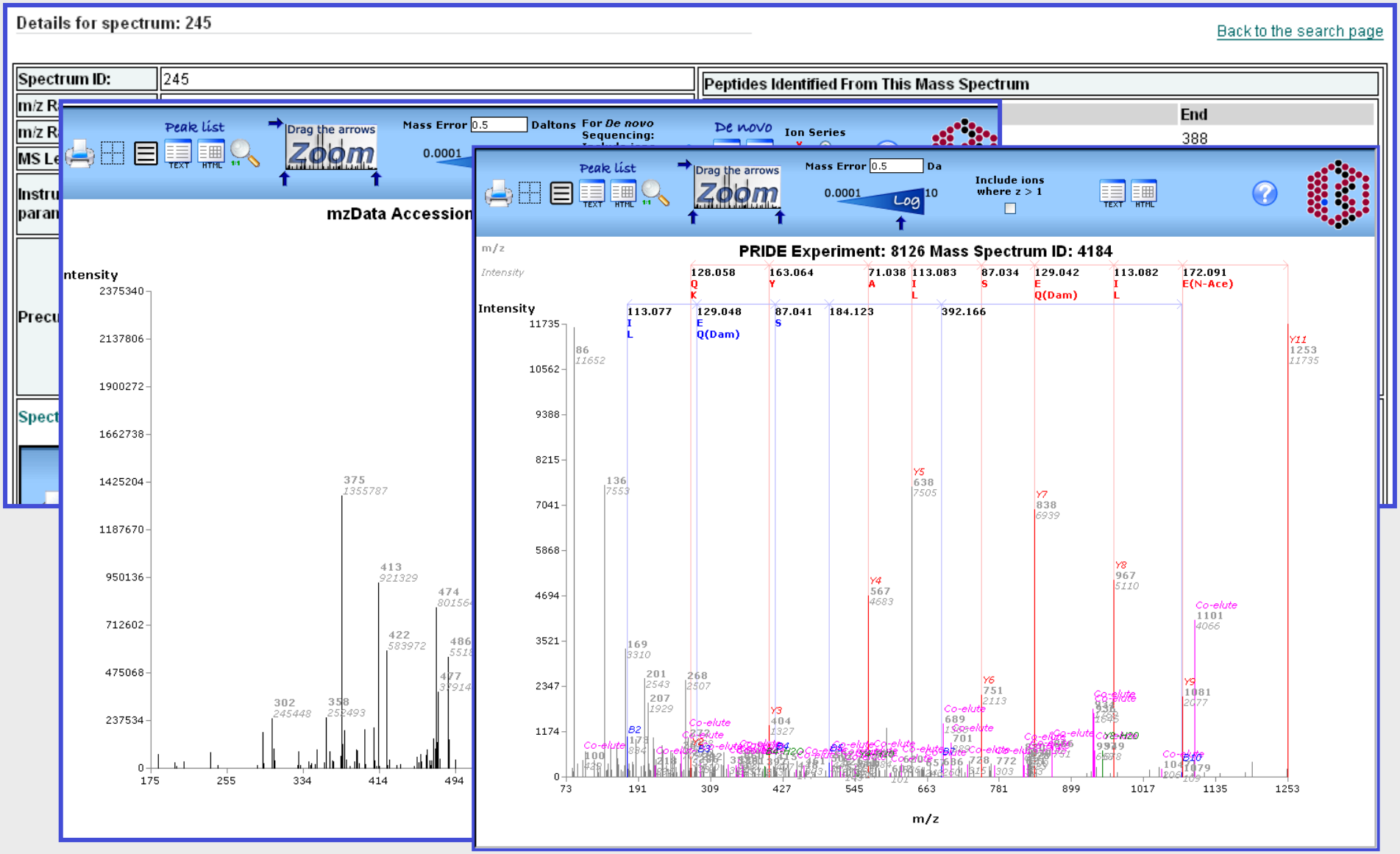
Sequence	Start	End
RELDESLQVAER	377	388

Spectrum View Spectrum Information

Additional

Source	Name	Value
PRIDE	PeptideProphet probability score	0.8438
PRIDE	X correlation	2.650
PRIDE	Delta Cn	0.189
PRIDE	Sp	997.9

PRIDE web interface – mass spectra



PRIDE BioMart

EMBL-EBI | EB-eye Search NEW | All Databases | Enter Text Here | Go | Reset | Advanced Search

Databases | Tools | Groups | Training | Industry | About Us | Help | Site Index

EBI > Databases > Proteomics > PRIDE > PRIDE BioMart

New | XML | Help | Count | Results

» **Dataset:**
PRIDE
» **Attributes**
Submitted Protein Accession
» **Filters**
Filter by Experiment Accession : [ID-list specified]

» **Dataset:**
[None Selected]

Display maximum 10 rows as HTML
Export all results to File Go

Submitted Protein Accession
15079369
Q15526
Q9Y6A2
21389381
17149828
O00273
21359982
P04901
P43251
Q9H5N1

biomart version 0.5

The spectacular bit: across-BioMart queries!

Question: "Which proteins, identified in PRIDE, in blood plasma, → PRIDE
are transcribed from genes located in chromosome 11" → Ensembl

biomart

HOME | MARTVIEW | MARTSERVICE | DOCS | CONTACT | NEWS | CREDITS

New | Count | Results | URL | XML | Perl | Help

Dataset 1895 / 37435 Genes
Homo sapiens genes (NCBI36)
Filters
Chromosome: 11
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Gene Start (bp)
Gene End (bp)
Chromosome Name
Associated Gene Name
Ensembl Protein ID

Dataset 498 / 8173 Experiments
PRIDE
Filters
Filter by Tissue : blood plasma
Attributes
PRIDE Experiment Accession
Experiment Title
Submitted Protein Accession
Uniprot Accession

Export all results to File TSV Unique results only Go
Email notification to

View 10 rows as HTML Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Gene Start (bp)	Gene End (bp)	Chromosome Name	Associated Gene Name	Ensembl Protein ID	PRIDE Experiment Accession	Experiment Title	Submitted Protein Accession	Uniprot Accession
ENSG00000221842	ENST00000335295	5203272	5207201	11	HBB	ENSP00000333994	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00218816	P68871
ENSG00000118137	ENST00000236850	116211677	116213571	11	APOA1	ENSP00000236850	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00021841	P02647
ENSG00000118137	ENST00000375320	116211677	116213571	11	APOA1	ENSP00000364469	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00021841	P02647
ENSG00000118137	ENST00000375323	116211677	116213571	11	APOA1	ENSP00000364472	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00021841	P02647
ENSG00000118137	ENST00000359492	116211677	116213571	11	APOA1	ENSP00000352471	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00021841	P02647
ENSG00000180210	ENST00000311907	46697331	46717631	11	F2	ENSP00000308541	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00019568	P00734
ENSG00000149131	ENST00000278407	57121436	57138902	11	SERPING1	ENSP00000278407	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00291866	P05155
ENSG00000110245	ENST00000375345	116205834	116208998	11	APOC3	ENSP00000364494	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00021857	P02656
ENSG00000110245	ENST00000227667	116205834	116208998	11	APOC3	ENSP00000227667	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00021857	P02656
ENSG00000110169	ENST00000265983	6408858	6418830	11	HPX	ENSP00000265983	13	HUPO Plasma Proteome Project, Lab # 2 Expt # 17	IPI00022488	P02790

biomart version 0.7

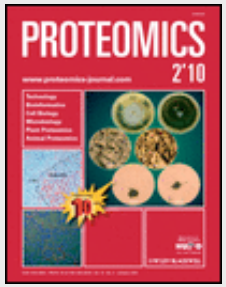
www.biomart.org

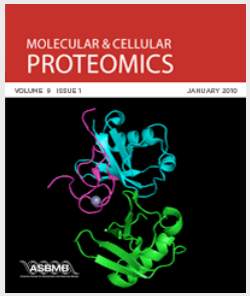
DATA SUBMISSION TO PRIDE

Journals recommend PRIDE as submission point

- Journal guidelines recommend now submission to proteomics repositories:

  - *Proteomics*
  - *Nature Biotechnology*
  - *Nature Methods*
  - *Molecular and Cellular Proteomics*





- Closer collaboration between *Proteomics* and PRIDE:

- "Deposition of supporting data in a public, open access database like PRIDE or World-2DPAGE is strongly recommended, and **mandatory** for Dataset Briefs"

MCP new guidelines

New guidelines from **MCP** for **data deposition**:

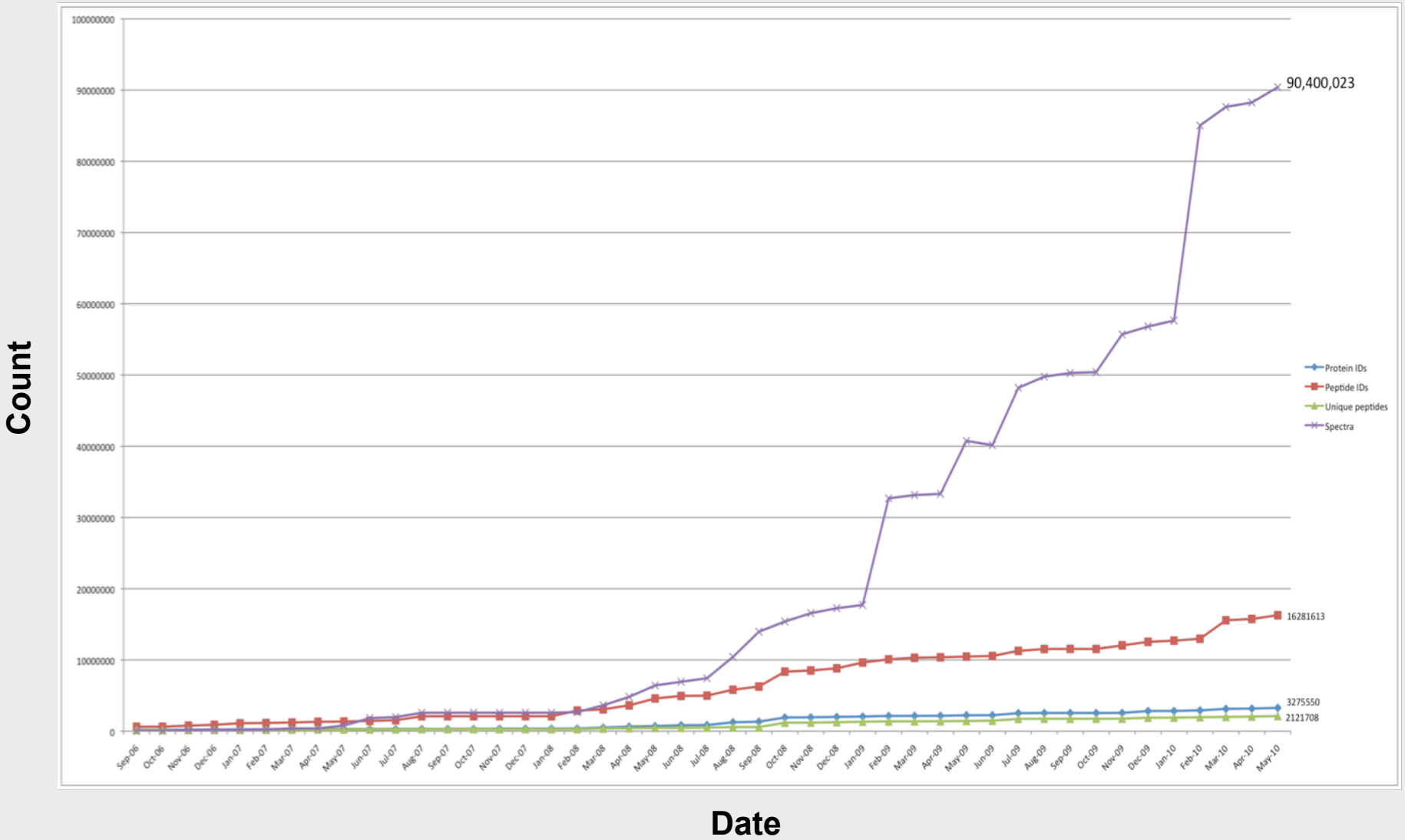
For all proteins identified on the basis of **ONE OR TWO unique peptide spectra**, the ability to view **annotated spectra** for these identifications must be made available. This can be achieved in one of three ways:

1) Submission of spectra and search results to a **public results repository** that is **equipped with a spectral viewer** (e.g. **PRIDE**, Peptidome etc). This information will appear as a **hyperlink** in the published article…

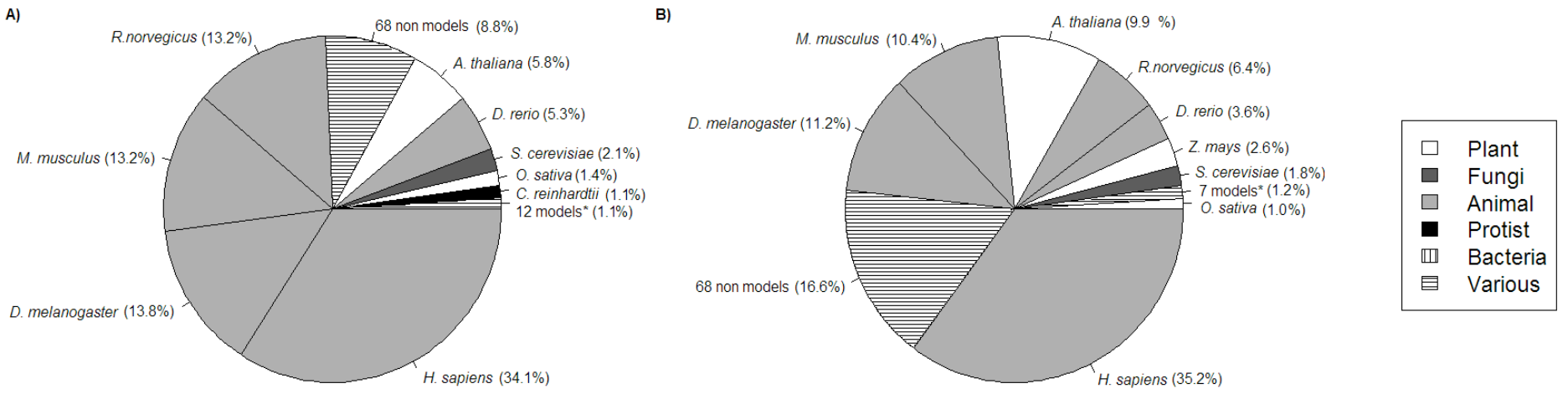
2) Submission (with the manuscript) of spectra and search results in a file format that allows visualization of the spectra using **a freely-available viewer.**

3) Submission (with the manuscript) of annotated spectra in an 'office' or PDF format.

PRIDE growth



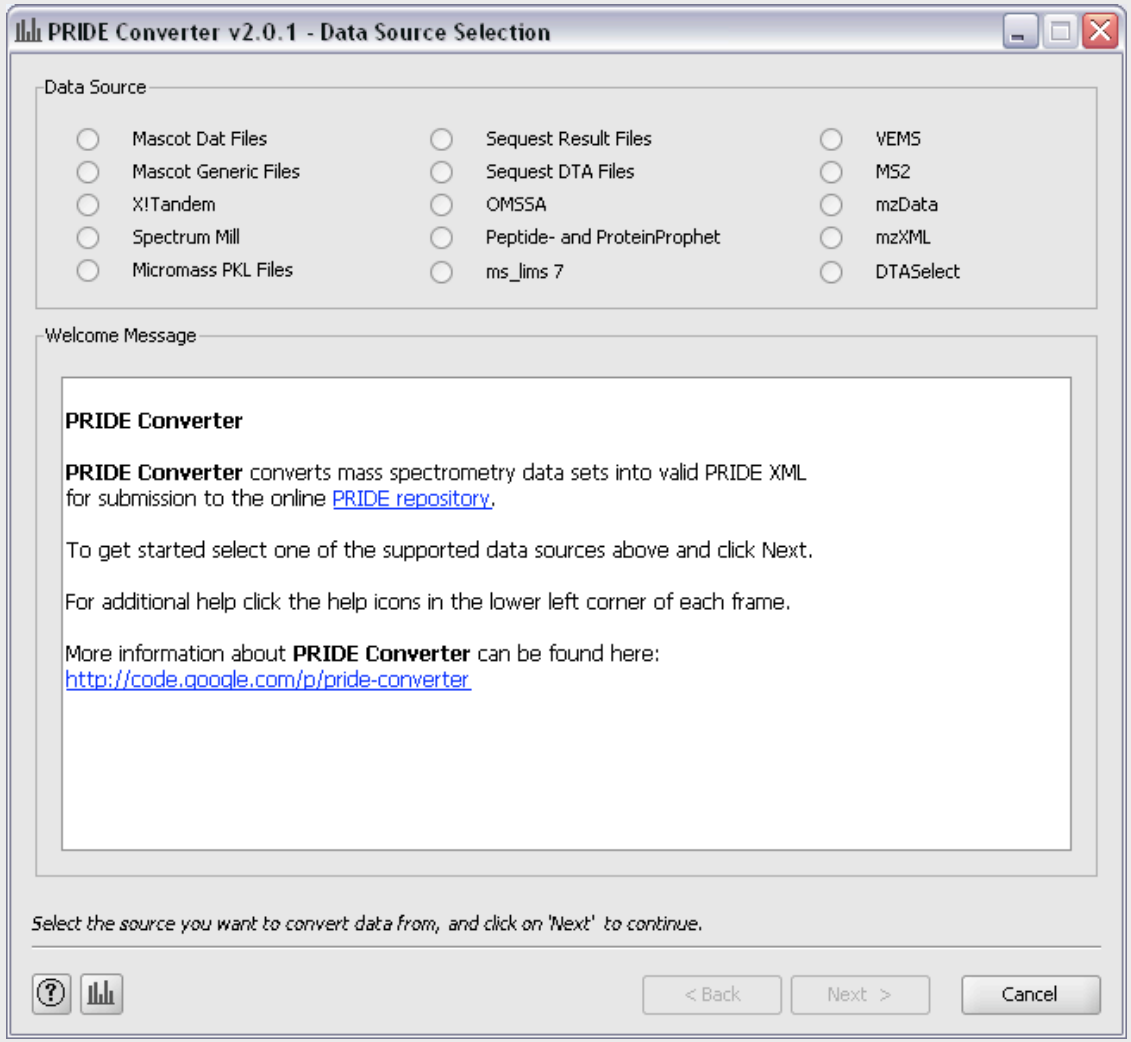
PRIDE data content



Protein IDs **Peptide IDs**

Why? Submission made easier: PRIDE Converter

http://code.google.com/p/pride-converter



Barsnes *et al.*, 2009

PRIDE Converter – interface details

PRIDE Converter v1.16.2 - Spectra Selection - Step 2 of 8

Simple Spectra Selection

Select All Spectra

Select Identified Spectra

Advanced Spectra Selection

Select Spectra Based On:

Manual Spectra Selection

PID Filename

Right click on a row to add spectru

Select the spectra to include in the PRID

PRIDE Converter v1

Samples Set

Name: Humar

Single Samp

Preferred O

1 Ho

Select a sample se

PRIDE Converter v1.16.2 - Protocol Properties - Step 5 of 8

Protocol

Name:

Protocol S

1 [R

2 [A

3 [E

Identification

Identificat

Select a proto

PRIDE Converter v1.16.2 - Instrument Properties - Step 6 of 8

Instrument

Name: Bruker Ultraflex

Source: Matrix-assisted Laser Desorption Ionization [PSI:1000075]

Detector: Electron Multiplier Tube [PSI:1000111]

Analyzers:

	CV Terms
1	[Bruker Daltonics ultraFlex TOF/TOF MS]

Add Analyzer

Processing

Software Name: FlexAnalysis

Software Version: 2.4

Processing Methods:

	CV Terms	Value
1	Deisotoping [PSI:1000033]	false
2	ChargeDeconvolution [PSI:1000034]	false
3	PeakProcessing [PSI:1000035]	CentroidMassSpectrum

Add Processing Method

Select an instrument from the list, or create your own, and click on 'Next' to continue.

< Back Next > Cancel

From PRIDE Converter to PRIDE FTP

PRIDE Converter v2.3.4 – Output Properties – Step 8 of 8

Output Details

Output Folder: /Users/javizca/

Resubmission

Resubmission * Original Accession Number:

* When resubmitting a PRIDE XML file please provide the original accession

Format Specific Parameters

Round Score and Threshold Down Before Comparison

Use Comma As

Transforming Spectra. Please

F007448.dat (1/1)

OMSSA Folder:

PRIDE Submission

File Created:

If the data is part of a paper, please include a reference to PRID

Select an output folder and click on 'Convert!' to create the PRIDE XML file.

Request PRIDE FTP Access?

Submission Tips

The size of your PRIDE XML file is larger than the maximum file size for using the 'Direct Submission' via the PRIDE web page.

We therefore recommend using the PRIDE FTP server. To get access to the FTP server, please contact the PRIDE team at pride-support@ebi.ac.uk.

Close

.xml.gz

cations (33.02 MB)

If the data is part of a paper, please include a reference to PRIDE Converter: PMID:19587657

PRIDE Login

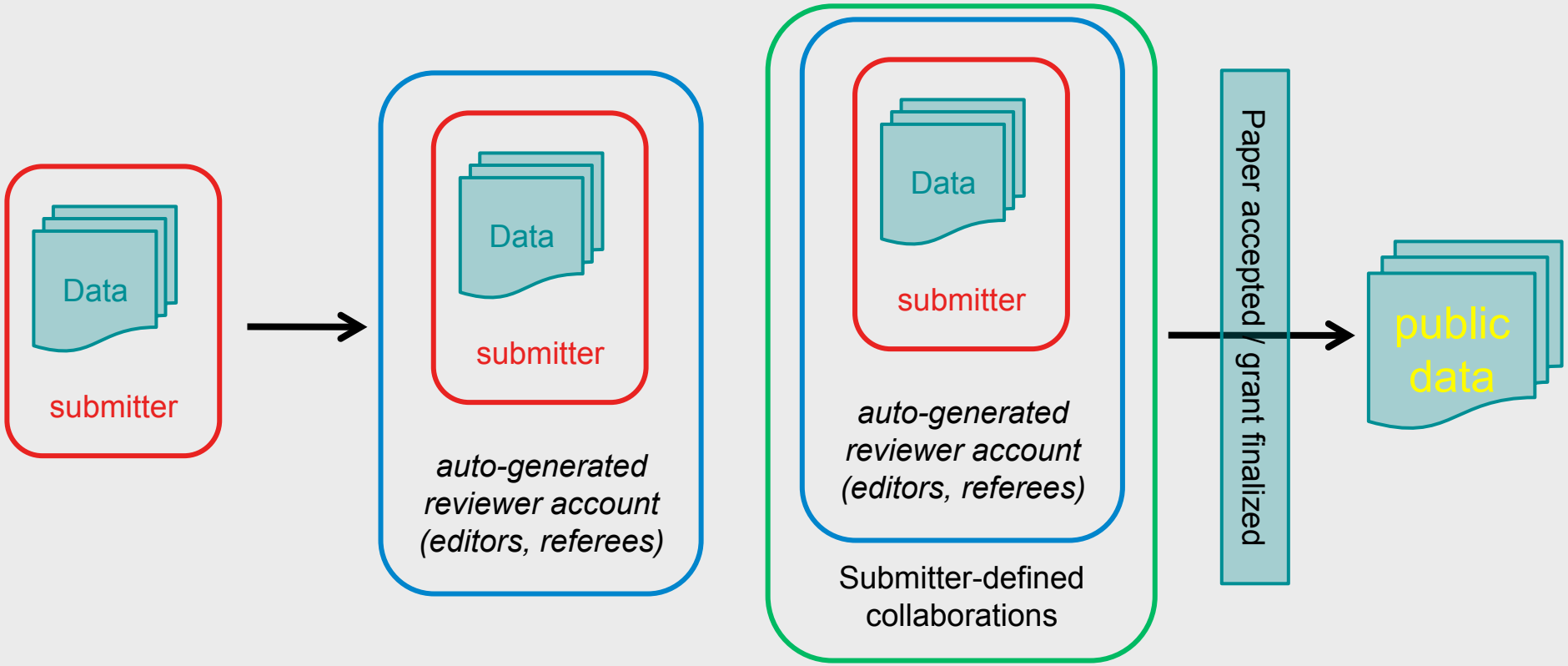
PRIDE Registration

< Back

Convert!

Exit

Data access privileges in PRIDE



PRIDE relies on a simple but very powerful group-based access system that can accommodate even more complex data release schemes than pictured here

OTHER PROTEOMICS REPOSITORIES

Existing proteomics repositories

- Main public repositories:
  - PROteomics IDEntifications database (PRIDE)
  - Global Proteome Machine (GPMDB)
  - Peptide Atlas
  - Tranche
  - NCBI Peptidome

- Smaller scale repositories, more specialized:
  Among others: Human Proteinpedia, Genome Annotation Proteomics Pipeline (GAPP), MAPU, SwedCAD, PepSeeker, Open Proteomics Database, …

- Very diverse: different aims, functionalities, …

Other MS proteomics repositories

PeptideAtlas	the gpmdb	PRIDE	peptidome NCBI Peptide Data Resource	Tranche
Reprocesses data	*Reprocesses data*	No reprocessing	No reprocessing	No reprocessing
Editorial control	*Editorial control*	No editorial control	No editorial control	No editorial control
Limited annotation	*Limited annotation*	Detailed annotation	Detailed annotation	*Limited annotation*
??	170 million peptides	96 million spectra	3.8 million spectra	??
??	22.3 million protein IDs	3.7 million protein IDs	60,000 protein IDs	??

PeptideAtlas

- Peptide identifications from MS/MS

- Data are **reprocessed** using the popular ***Trans Proteomic Pipeline*** *(TPP)*

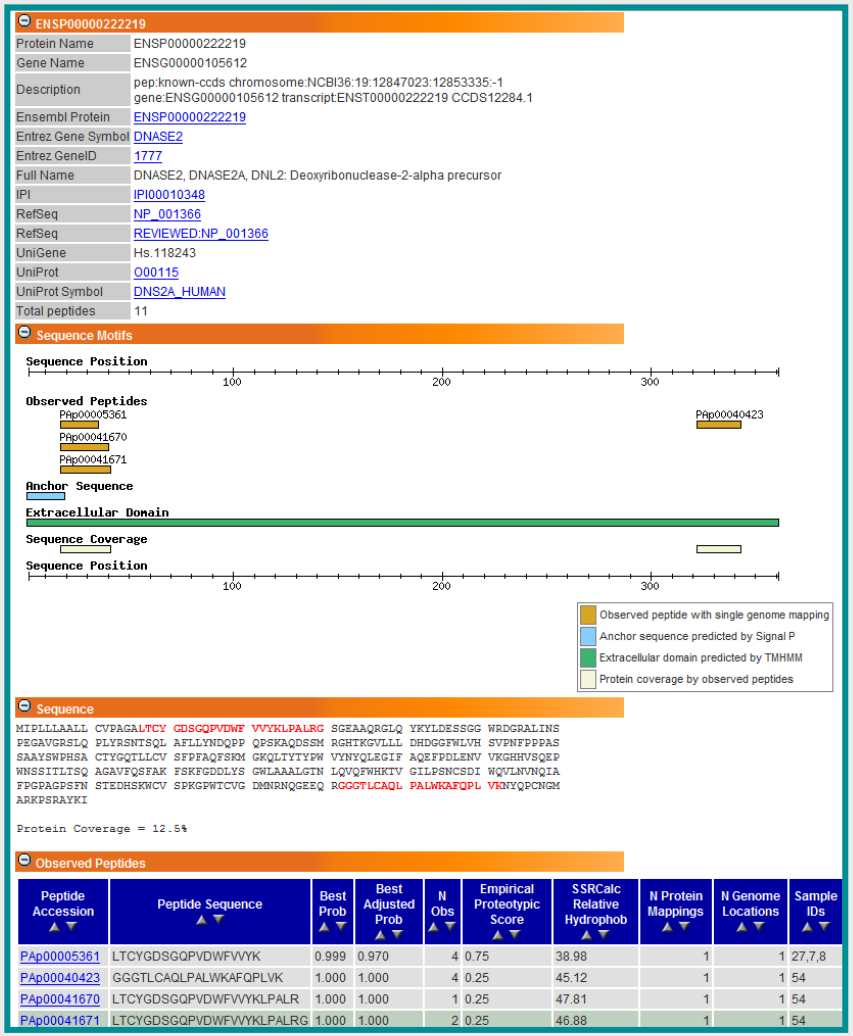
- Uses ***PeptideProphet*** to derive a probability for the correct identification for all contained peptides

http://www.peptideatlas.org

PeptideAtlas

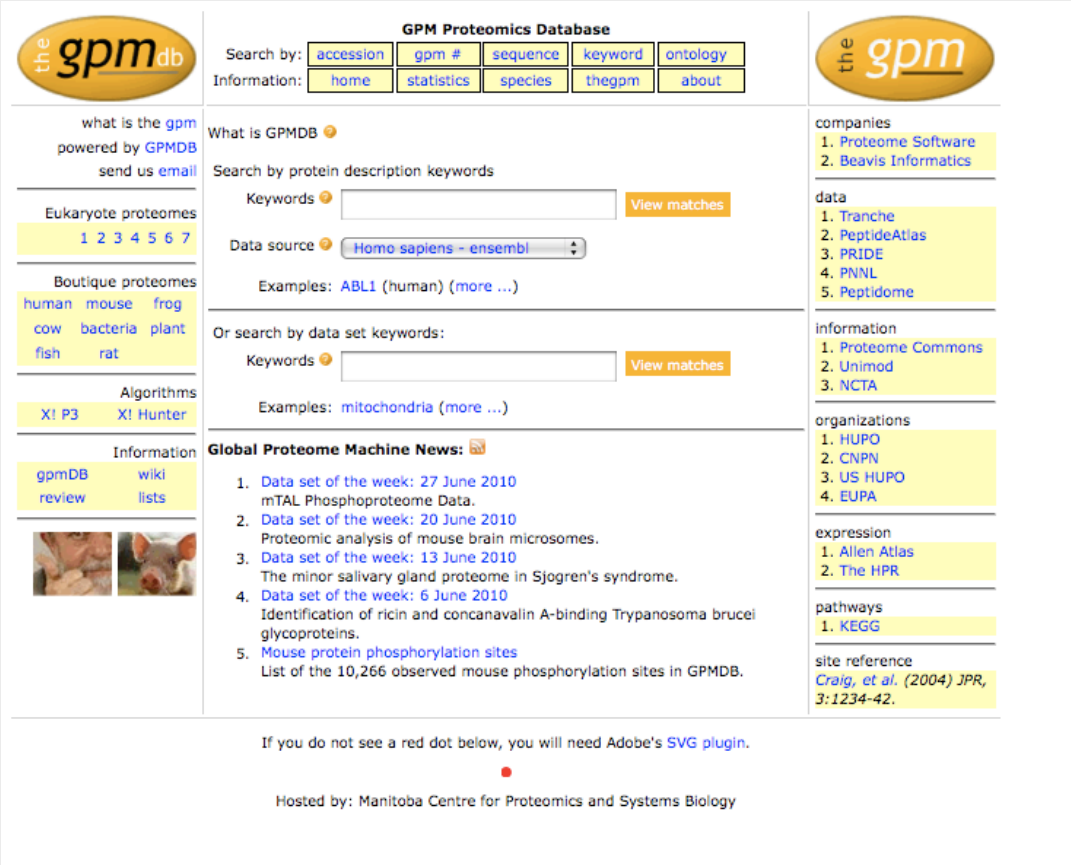
- All peptides mapped to ***Ensembl*** using ***ProteinProphet*** *(for human)*
- Built by the Aebersold lab to help them find proteotypic peptides
- Provides proteotypic peptide predictions
- Limited metadata
- Great support for **targeted proteomics approaches** (SRM/MRM)

http://www.peptideatlas.org



GPMDB

- End point of the *GPM proteomics pipeline*, to aid in the process of validating peptide MS/MS spectra and protein coverage patterns.



http://gpmdb.thegpm.org/

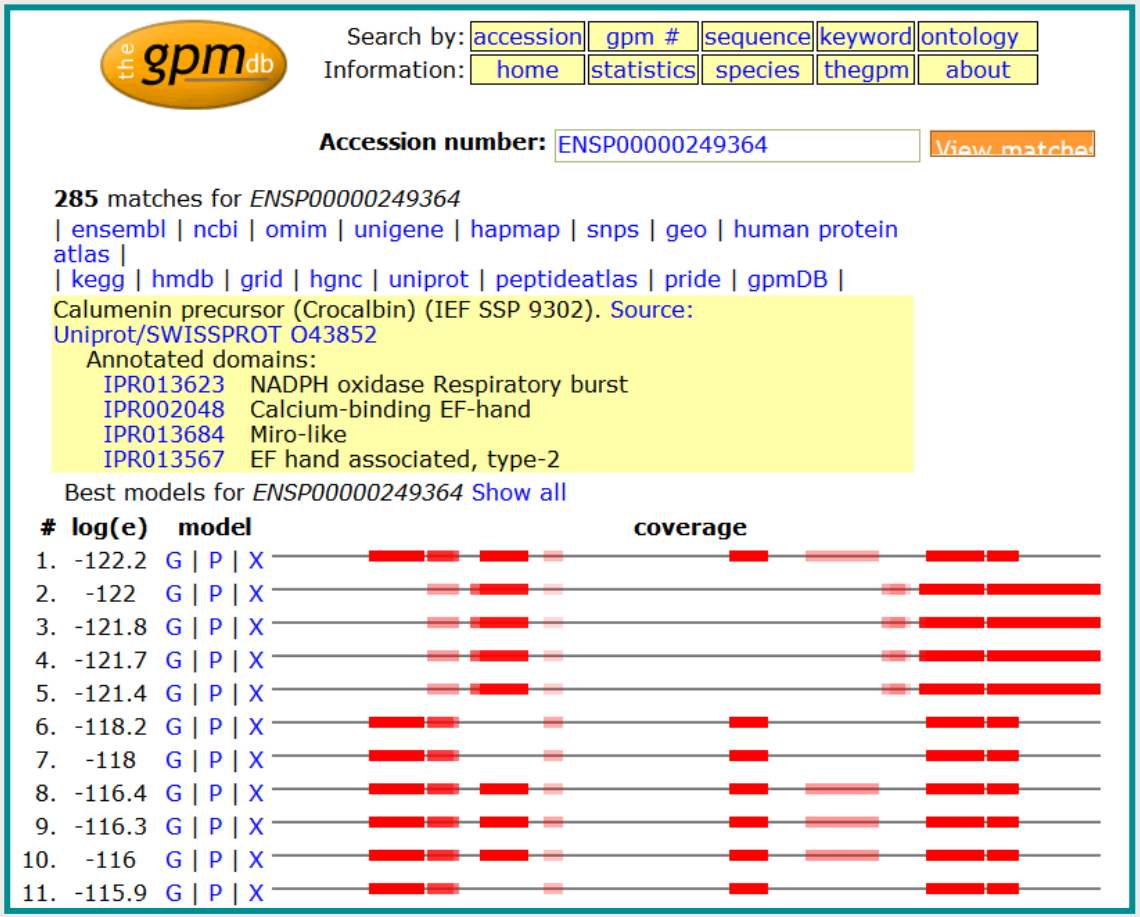
GPMDB

- End point of the *GPM proteomics pipeline*, to aid in the process of validating peptide MS/MS spectra and protein coverage patterns.
- Data are **reprocessed** using the popular ***X!Tandem*** *or* ***X! Hunter*** spectral searching algorithm
- Also provides proteotypic peptides

http://gpmdb.thegpm.org/

GPMDB

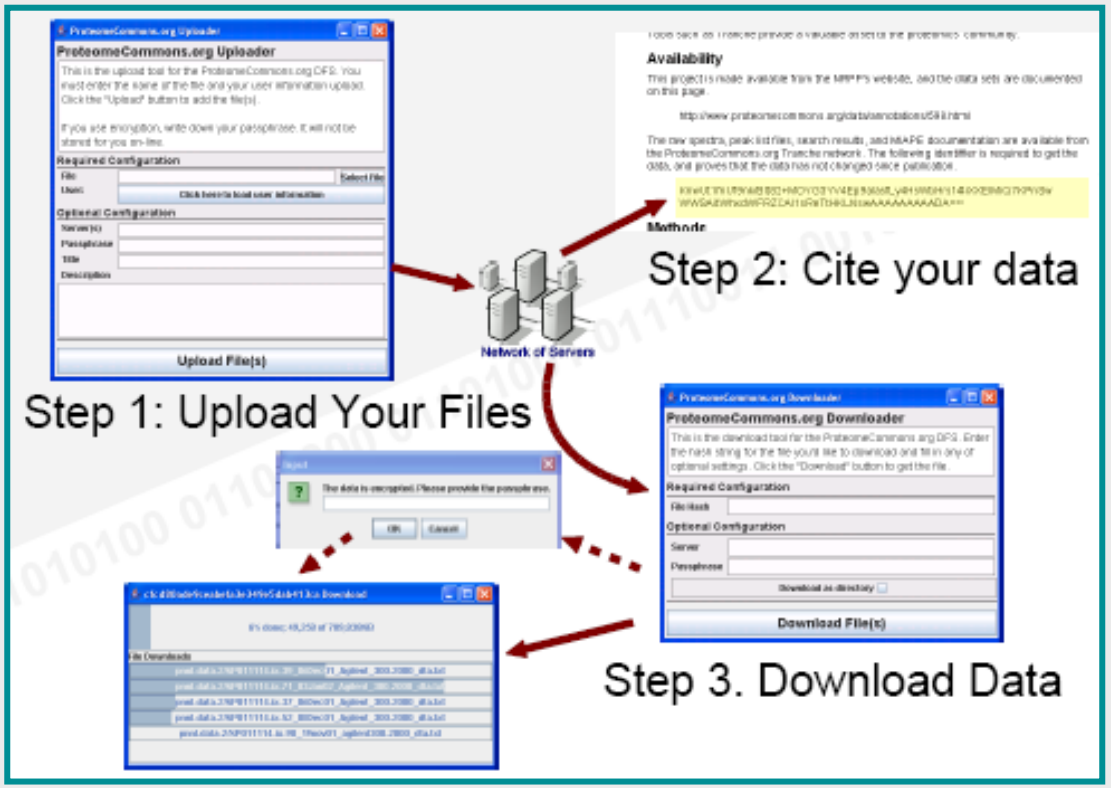
- Powerful visualization features
- Provides very limited annotation with GO, BTO
- Some support to targeted approaches is available



http://gpmdb.thegpm.org/

Tranche

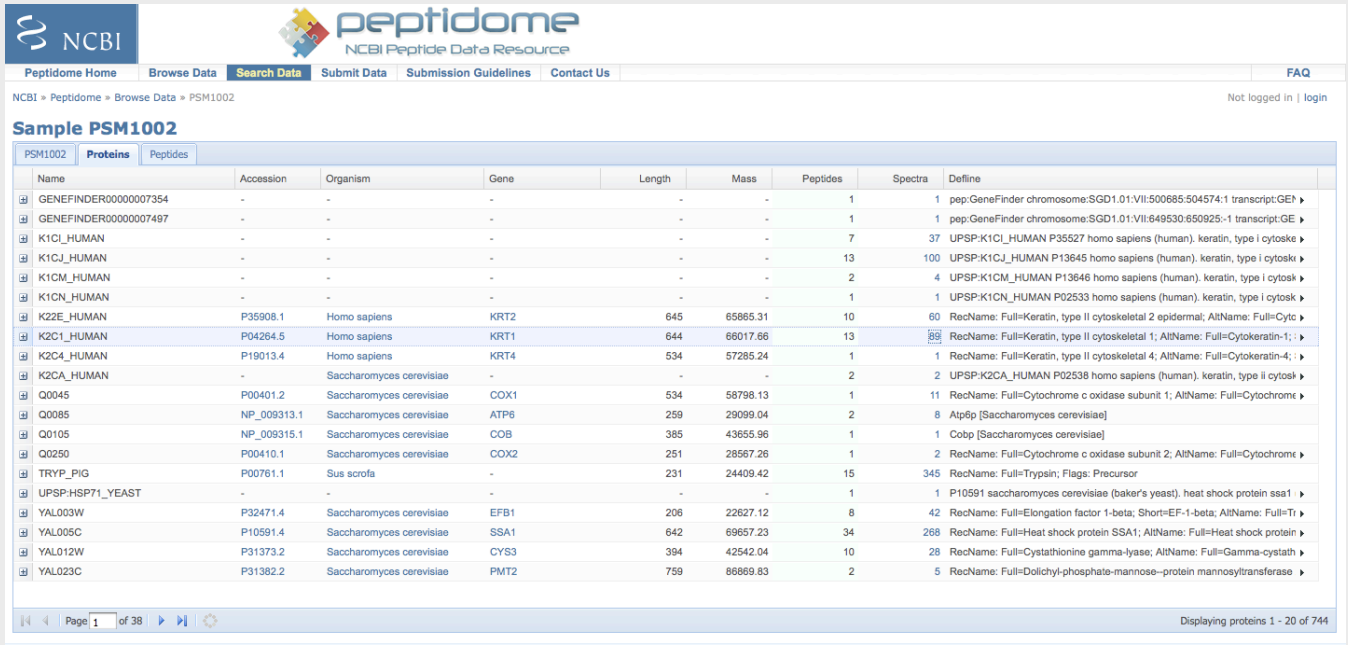
- Peer-to-peer distributed filesystem (original name: *the DFS*)
- Meant to securely store, and conveniently deliver large amounts of data
- Provides a highly specialized, but much needed niche service
- Has already been used by PRIDE to store certain large files
- Very limited annotation (metadata is not mandatory)



http://tranche.proteomecommons.org

NCBI Peptidome

- No reprocessing
- Detailed annotation (no CVs)
- Review system
- Sibling resource to PRIDE



Name	Accession	Organism	Gene	Length	Mass	Peptides	Spectra	Defline
GENEFINDER00000007354	-	-	-	-	-	1	1	pep:GeneFinder chromosome:SGD1.01:VII:500685:504574:1 transcript:GEF
GENEFINDER00000007497	-	-	-	-	-	1	1	pep:GeneFinder chromosome:SGD1.01:VII:649530:650925:-1 transcript:GE
K1CI_HUMAN	-	-	-	-	-	7	37	UPSP:K1CI_HUMAN P35527 homo sapiens (human). keratin, type i cytoske
K1CJ_HUMAN	-	-	-	-	-	13	100	UPSP:K1CJ_HUMAN P13645 homo sapiens (human). keratin, type i cytoski
K1CM_HUMAN	-	-	-	-	-	2	4	UPSP:K1CM_HUMAN P13646 homo sapiens (human). keratin, type i cytosk
K1CN_HUMAN	-	-	-	-	-	1	1	UPSP:K1CN_HUMAN P02533 homo sapiens (human). keratin, type i cytosk
K22E_HUMAN	P35908.1	Homo sapiens	KRT2	645	65865.31	10	60	RecName: Full=Keratin, type II cytoskeletal 2 epidermal; AltName: Full=Cytc
K2C1_HUMAN	P04264.5	Homo sapiens	KRT1	644	66017.66	13	89	RecName: Full=Keratin, type II cytoskeletal 1; AltName: Full=Cytokeratin-1;
K2C4_HUMAN	P19013.4	Homo sapiens	KRT4	534	57285.24	1	1	RecName: Full=Keratin, type II cytoskeletal 4; AltName: Full=Cytokeratin-4;
K2CA_HUMAN	-	Saccharomyces cerevisiae	-	-	-	2	2	UPSP:K2CA_HUMAN P02538 homo sapiens (human). keratin, type ii cytosk
Q0045	P00401.2	Saccharomyces cerevisiae	COX1	534	58798.13	1	11	RecName: Full=Cytochrome c oxidase subunit 1; AltName: Full=Cytochrome
Q0085	NP_009313.1	Saccharomyces cerevisiae	ATP6	259	29099.04	2	8	Atp6p [Saccharomyces cerevisiae]
Q0105	NP_009315.1	Saccharomyces cerevisiae	COB	385	43655.96	1	1	Cobp [Saccharomyces cerevisiae]
Q0250	P00410.1	Saccharomyces cerevisiae	COX2	251	28567.26	1	2	RecName: Full=Cytochrome c oxidase subunit 2; AltName: Full=Cytochrome
TRYP_PIG	P00761.1	Sus scrofa	-	231	24409.42	15	345	RecName: Full=Trypsin; Flags: Precursor
UPSP:HSP71_YEAST	-	-	-	-	-	1	1	P10591 saccharomyces cerevisiae (baker's yeast). heat shock protein ssa1
YAL003W	P32471.4	Saccharomyces cerevisiae	EFB1	206	22627.12	8	42	RecName: Full=Elongation factor 1-beta; Short=EF-1-beta; AltName: Full=Tr
YAL005C	P10591.4	Saccharomyces cerevisiae	SSA1	642	69657.23	34	268	RecName: Full=Heat shock protein SSA1; AltName: Full=Heat shock protein
YAL012W	P31373.2	Saccharomyces cerevisiae	CYS3	394	42542.04	10	28	RecName: Full=Cystathionine gamma-lyase; AltName: Full=Gamma-cystath
YAL023C	P31382.2	Saccharomyces cerevisiae	PMT2	759	86869.83	2	5	RecName: Full=Dolichyl-phosphate-mannose--protein mannosyltransferase

http://www.ncbi.nlm.nih.gov/peptidome/

Other MS proteomics repositories

PeptideAtlas	the gpmdb	PRIDE	peptidome NCBI Peptide Data Resource	Tranche
Reprocesses data	*Reprocesses data*	No reprocessing	No reprocessing	No reprocessing
Editorial control	*Editorial control*	No editorial control	No editorial control	No editorial control
Limited annotation	*Limited annotation*	Detailed annotation	Detailed annotation	*Limited annotation*
??	162 million peptides	92 million spectra	3.8 million spectra	??
??	21.5 million protein IDs	3.5 million protein IDs	60,000 protein IDs	??

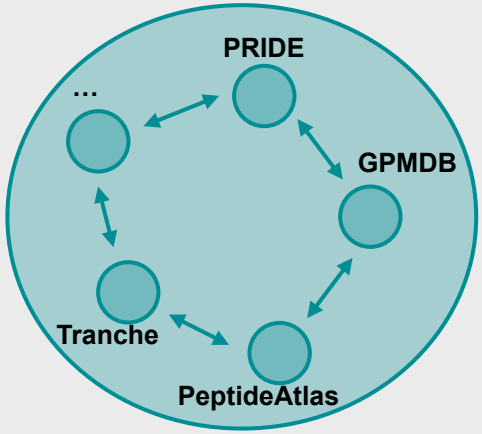
PRIDE AND OTHER REPOSITORIES: ProteomeXchange

Juan A. Vizcaíno
juan@ebi.ac.uk

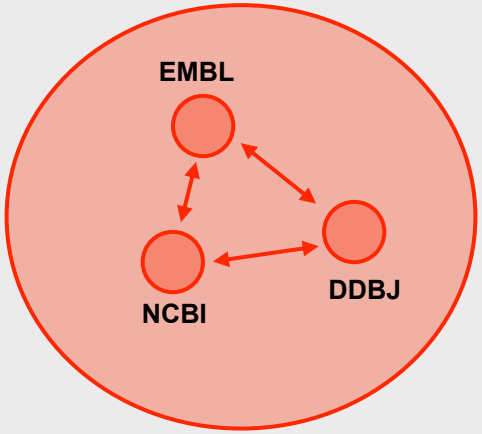
For sharing, superstructures must be built

Often, multiple repositories will emerge more or less simultaneously in a particular field. By exchanging data, and by collaborating on data acquisition an increase in coverage as well as a more comprehensive dataset is obtained by each individual resource.

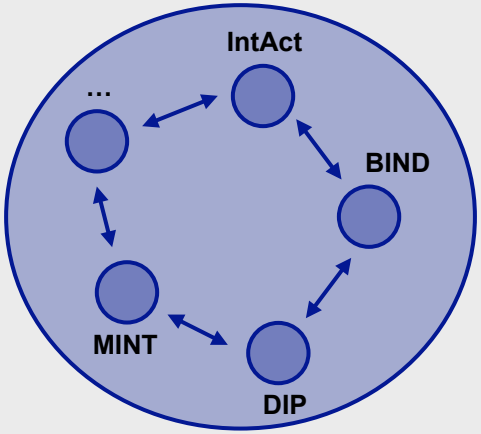
Such superstructures do require additional infrastructure, however.



mass spec
ProteomeXchange



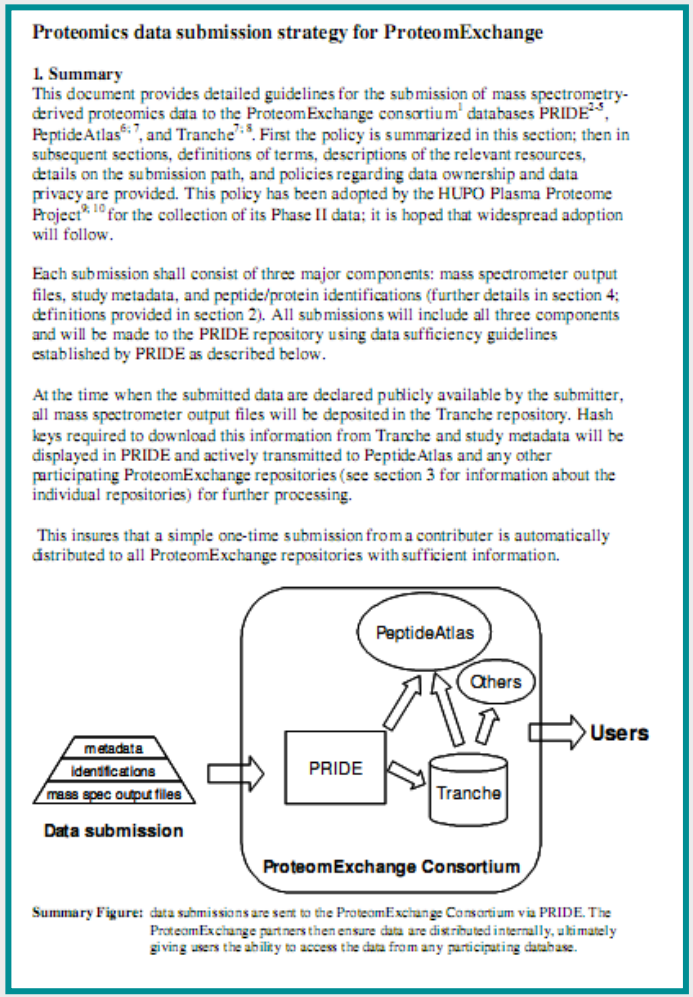
sequence databases
(INSDC)



interactions
IMEx

ProteomeXchange consortium

- Sharing proteomics data between existing proteomics repositories
- Includes PeptideAtlas, GPMDB, NCBI Peptidome and PRIDE, with data sharing infrastructure provided by Tranche
- Submission guidelines document finalized, it was proven on three different datasets
- ProteomeXchange is primarily **user-oriented**: the idea is to provide a **single point of submission**, but **multiple points of data visualization and analysis**



Proteomics data submission strategy for ProteomExchange

1. Summary

This document provides detailed guidelines for the submission of mass spectrometry-derived proteomics data to the ProteomExchange consortium[1] databases PRIDE[2-5], PeptideAtlas[6; 7], and Tranche[7; 8]. First the policy is summarized in this section; then in subsequent sections, definitions of terms, descriptions of the relevant resources, details on the submission path, and policies regarding data ownership and data privacy are provided. This policy has been adopted by the HUPO Plasma Proteome Project[9; 10] for the collection of its Phase II data; it is hoped that widespread adoption will follow.

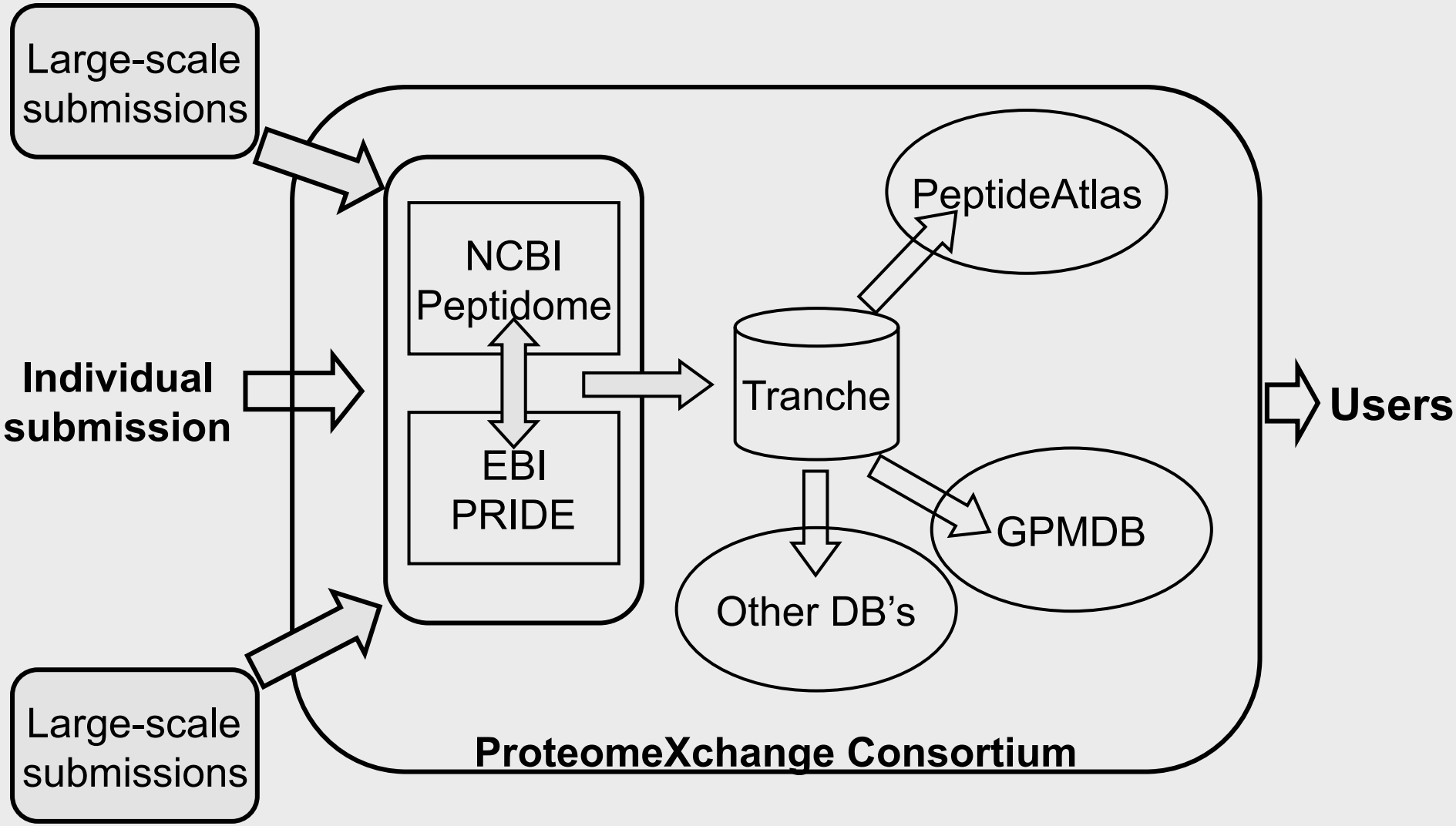
Each submission shall consist of three major components: mass spectrometer output files, study metadata, and peptide/protein identifications (further details in section 4; definitions provided in section 2). All submissions will include all three components and will be made to the PRIDE repository using data sufficiency guidelines established by PRIDE as described below.

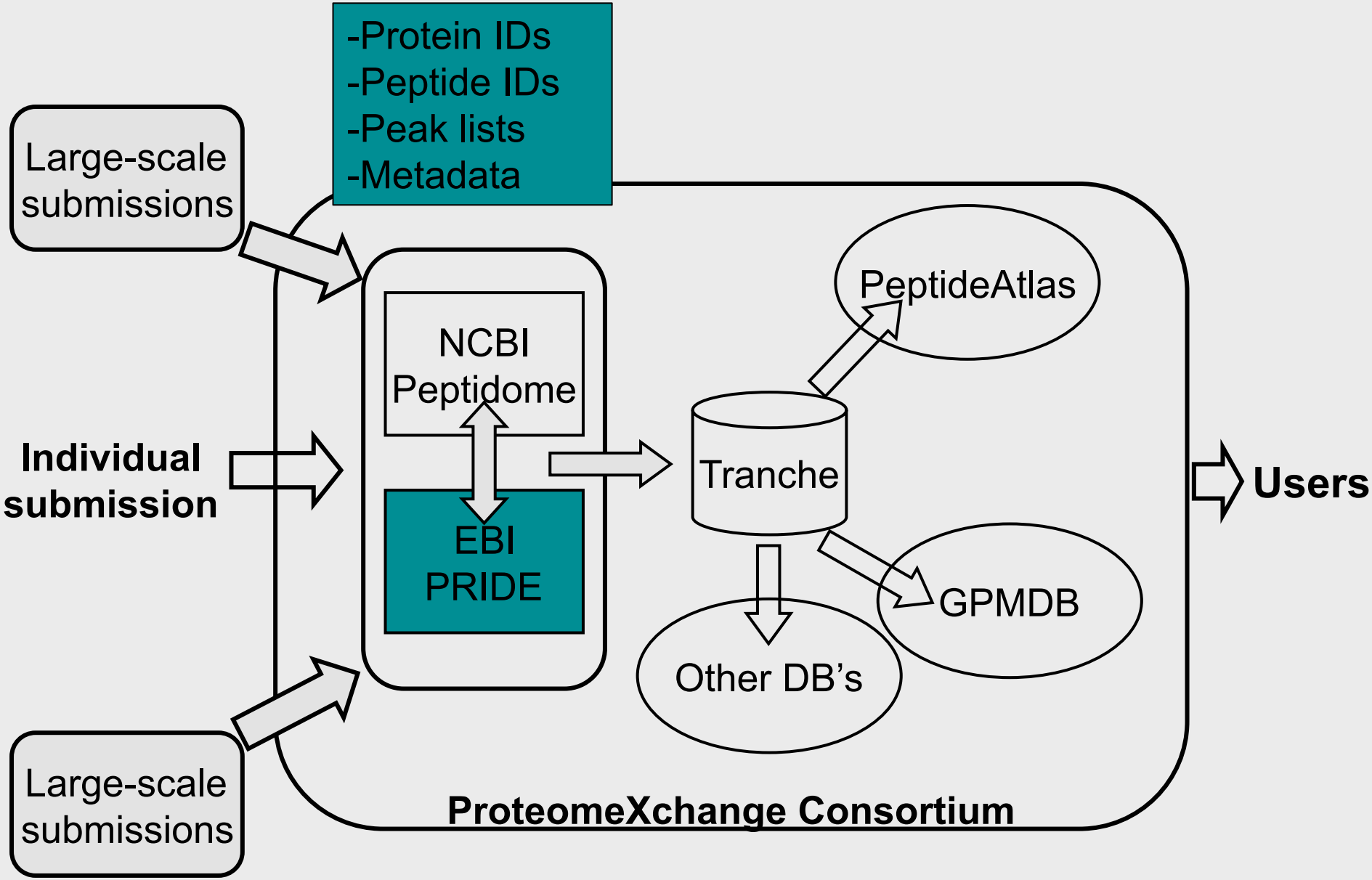
At the time when the submitted data are declared publicly available by the submitter, all mass spectrometer output files will be deposited in the Tranche repository. Hash keys required to download this information from Tranche and study metadata will be displayed in PRIDE and actively transmitted to PeptideAtlas and any other participating ProteomExchange repositories (see section 3 for information about the individual repositories) for further processing.

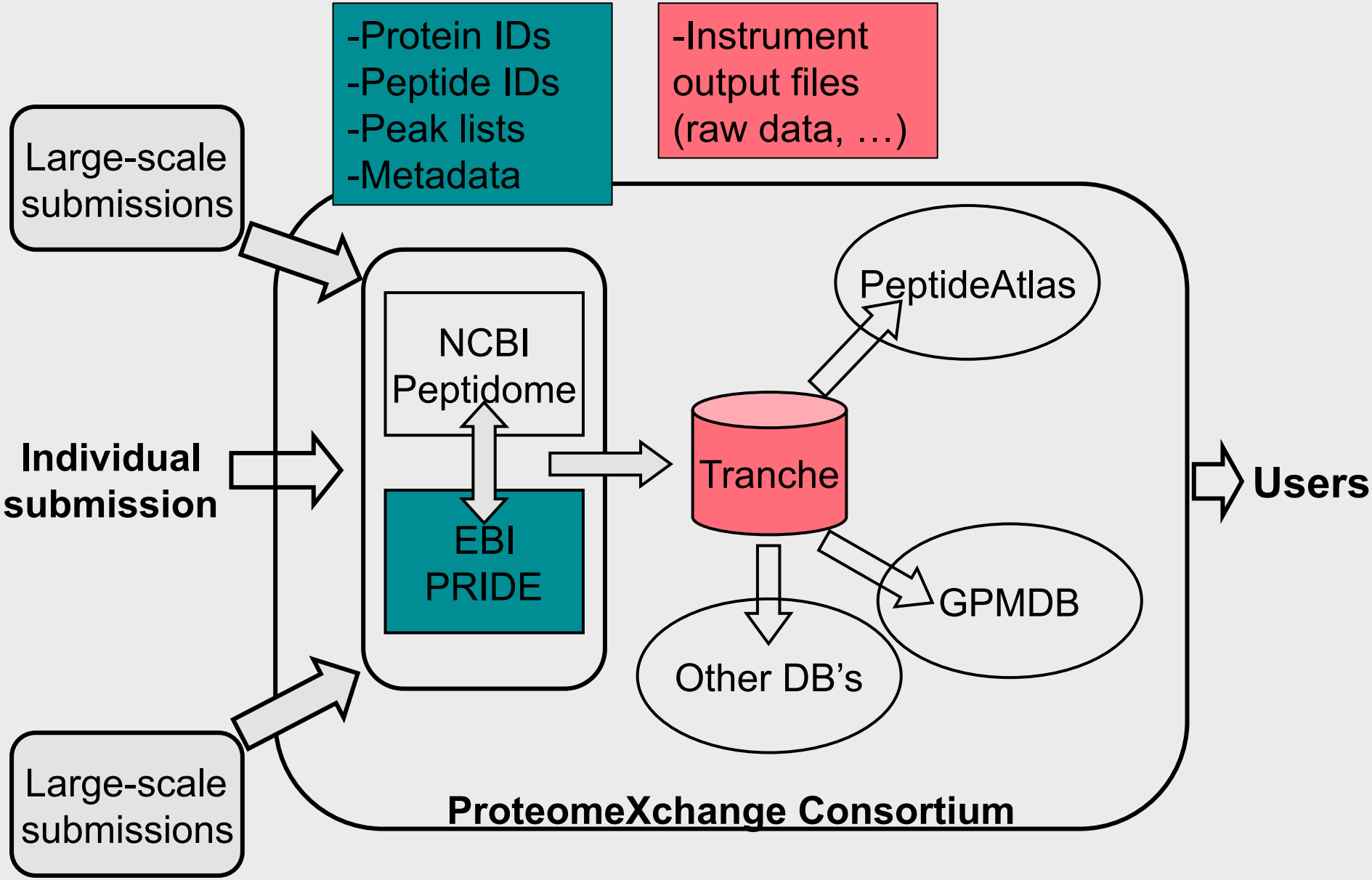
This insures that a simple one-time submission from a contributer is automatically distributed to all ProteomExchange repositories with sufficient information.

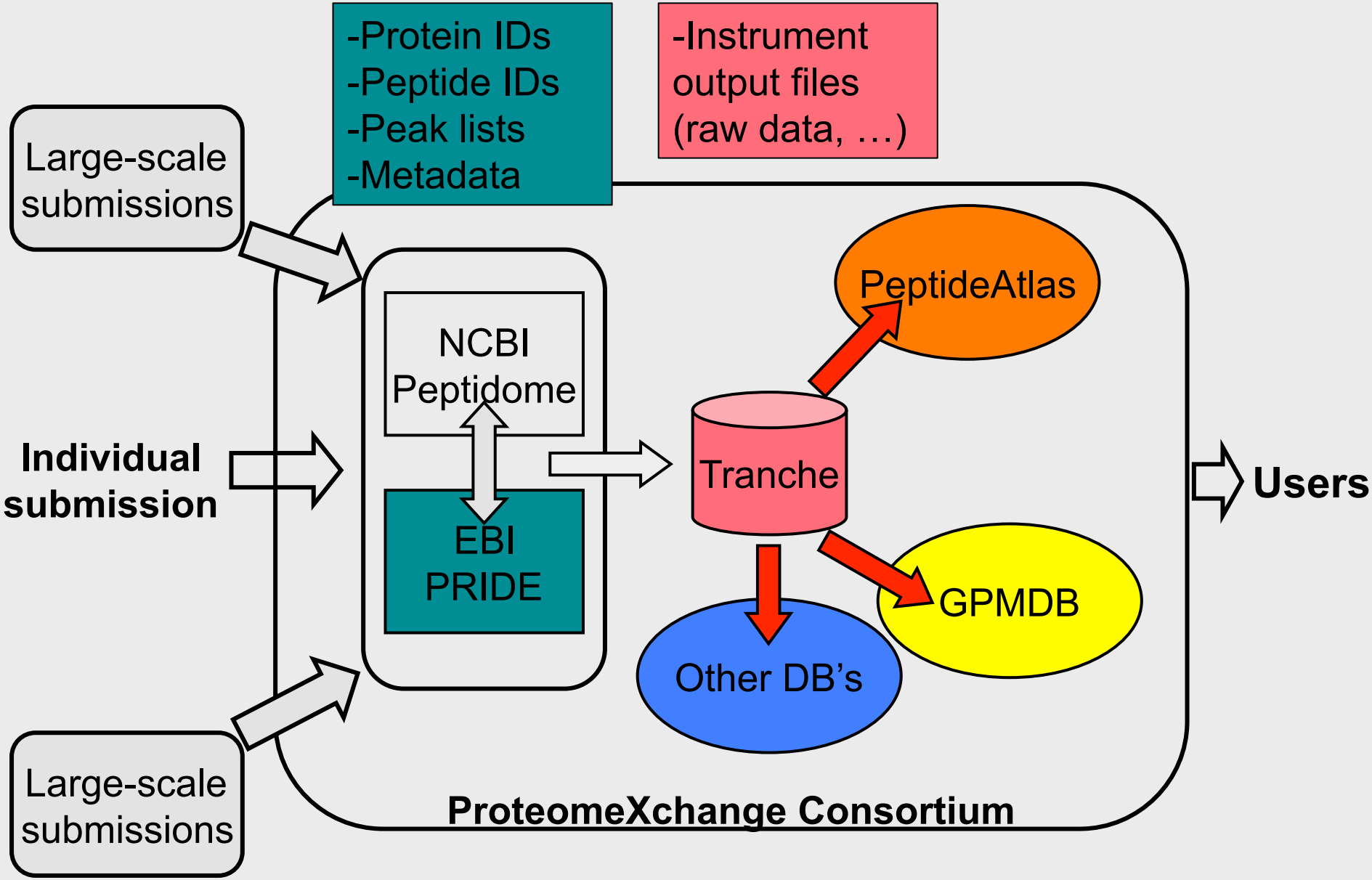
Summary Figure: data submissions are sent to the ProteomExchange Consortium via PRIDE. The ProteomExchange partners then ensure data are distributed internally, ultimately giving users the ability to access the data from any participating database.

www.proteomexchange.org


Large-scale submissions
Individual submission
Large-scale submissions
NCBI Peptidome
EBI PRIDE
Tranche
PeptideAtlas
GPMDB
Other DB's
Users
ProteomeXchange Consortium

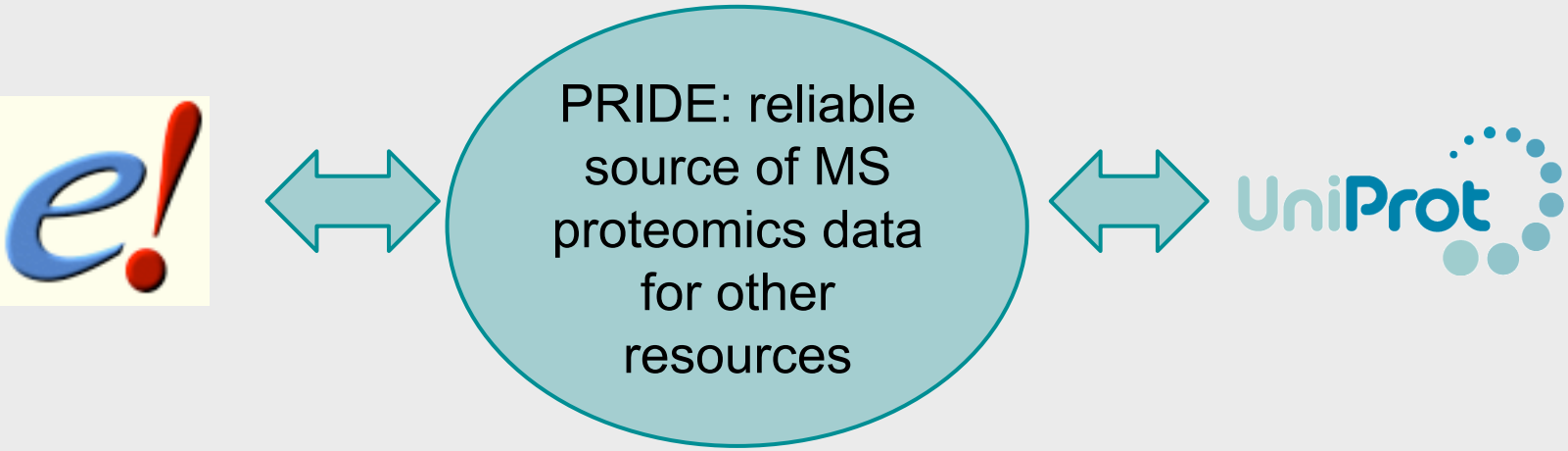

-Protein IDs
-Peptide IDs
-Peak lists
-Metadata
Large-scale submissions
Individual submission
Large-scale submissions
NCBI Peptidome
EBI PRIDE
Tranche
PeptideAtlas
GPMDB
Other DB's
Users
ProteomeXchange Consortium


-Protein IDs
-Peptide IDs
-Peak lists
-Metadata
-Instrument output files (raw data, …)
Large-scale submissions
Individual submission
Large-scale submissions
NCBI Peptidome
EBI PRIDE
Tranche
PeptideAtlas
GPMDB
Other DB's
Users
ProteomeXchange Consortium

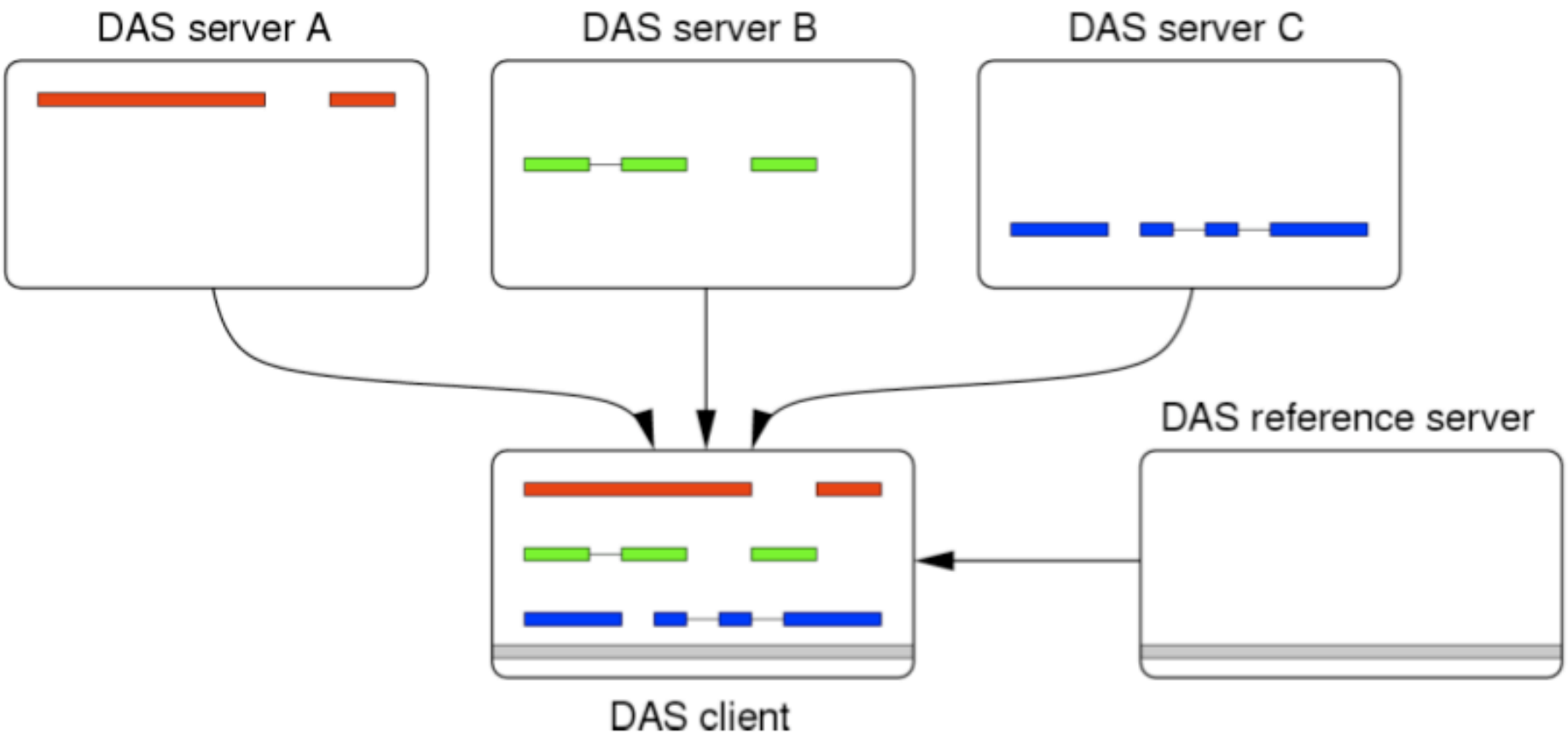

-Protein IDs
-Peptide IDs
-Peak lists
-Metadata
-Instrument output files (raw data, …)
Large-scale submissions
Individual submission
Large-scale submissions
NCBI Peptidome
EBI PRIDE
Tranche
PeptideAtlas
GPMDB
Other DB's
Users
ProteomeXchange Consortium

PRIDE: why is it there?

- Repository to support publications (proteomics MS derived data)
- **Source of proteomics data for other data resources**

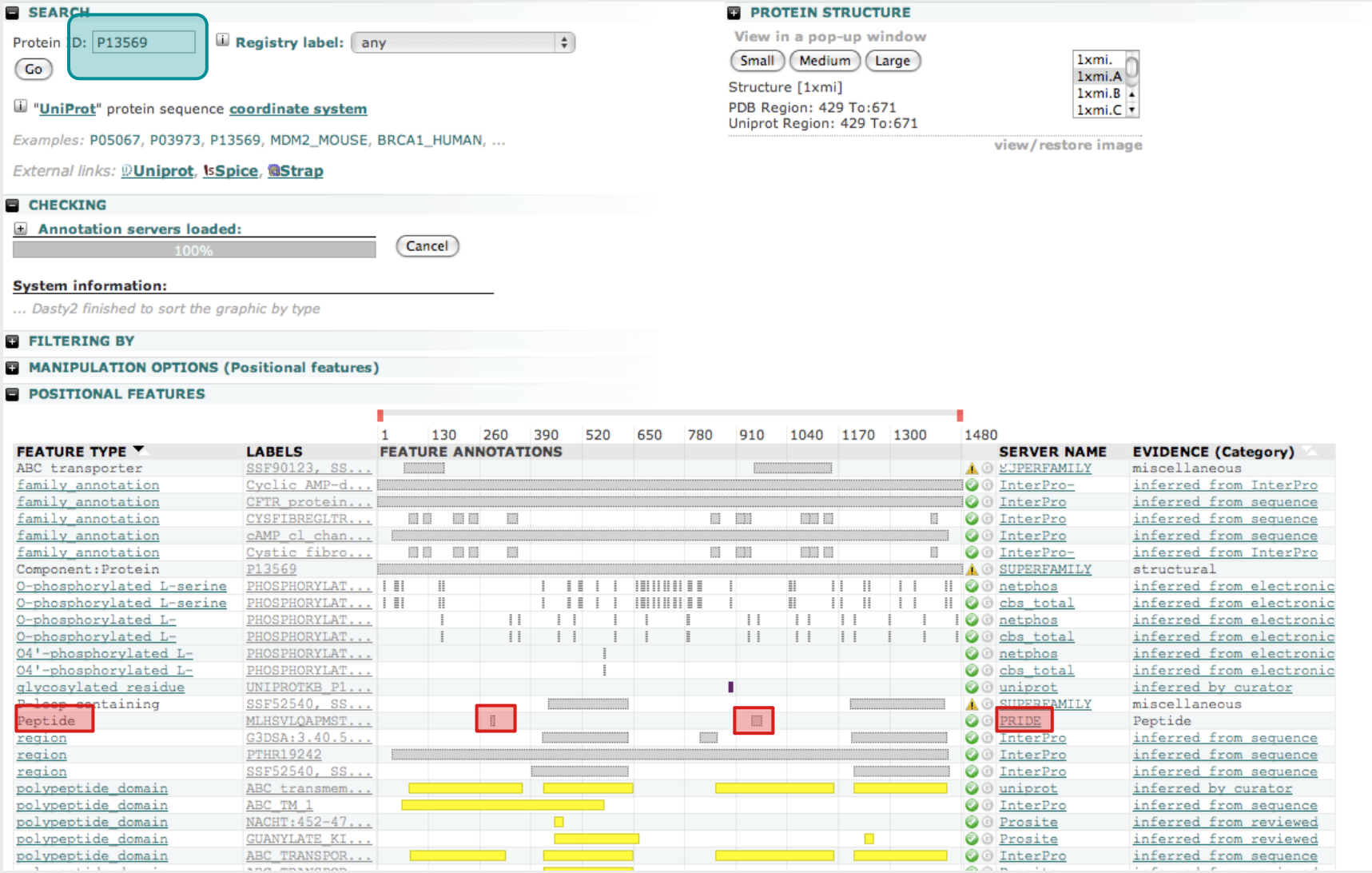


Distributed Annotation System

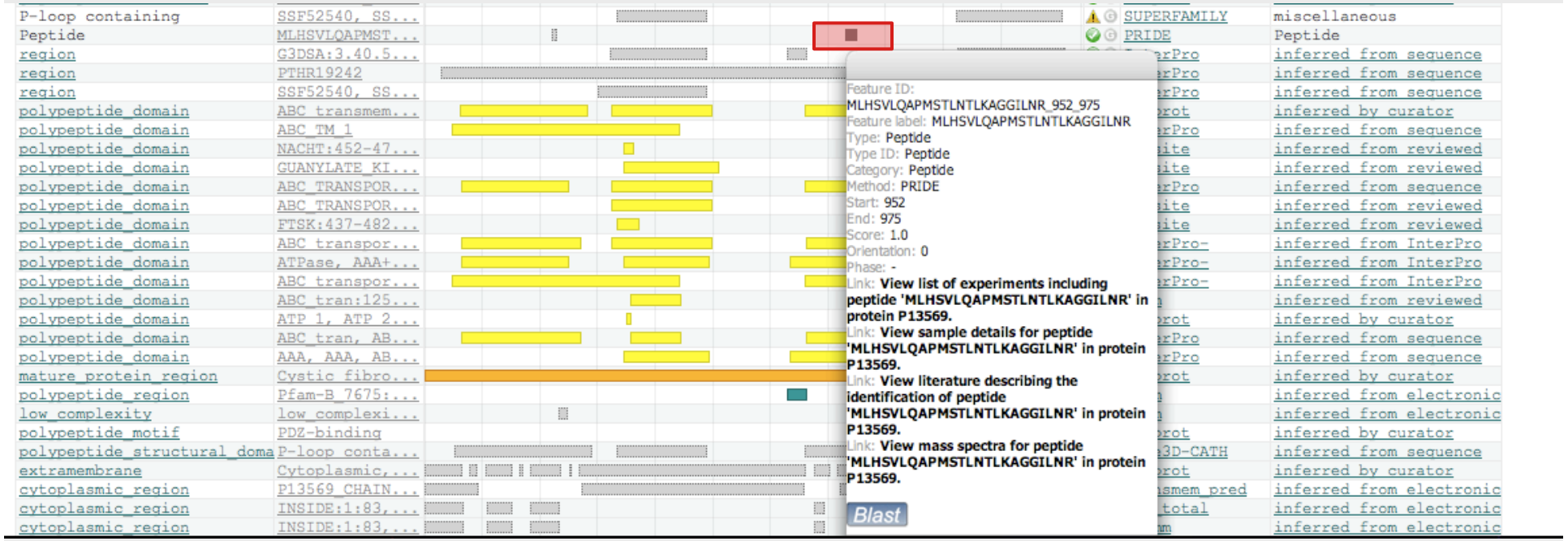


Dasty2 Version 0.5.1 - August 09 (http://www.ebi.ac.uk/dasty/)

PRIDE DAS server: Dasty example (1)



PRIDE DAS server: Dasty example (2)



Data sharing requires proper infrastructure

- Community supported, standardized data formats
  Necessary to allow efficient access to the data

- Controlled vocabularies (CVs) and ontologies
  To provide unambiguous context and metadata to the actual data, as well as enabling powerful queries to be performed on the data

- Minimal reporting requirements for specific data types
  Ensures the presence of certain bits of information without which interpretation is ambiguous, hampered or impossible

- Publicly available, online repositories
  Bioinformatics grew up along side the internet, and this is reflected in the successful online data sharing mechanism already in place in the life sciences. *The repositories should implement the standards, use the CV's and ontologies, and adhere to the minimal requirements.*

Coming soon… support for quantitative data

Details for identification: IPI00068506.1

Submitted Accession	IPI00068506.1
Search Database	IPI_HUMAN
Cross-References	

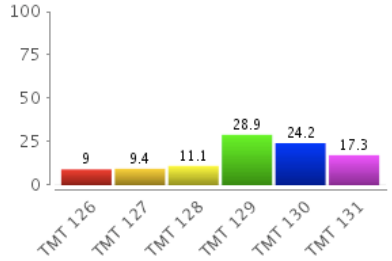
Accession	Database
ENST00000222388	ENSEMBL
ENSP00000222388	ENSEMBL_HUMAN
IPI00068506.1	IPI
NP_005683.2	REFSEQ
Q75MJ1.1	TREMBL
Q75MJ1_HUMAN	TREMBL
Q9UG63.1	TREMBL

These mappings have been obtained using the Protein Identifier Cross Reference (PICR) Service at the EBI. They are based on 100% sequence identity and, as a further requirement where applicable, all submitted peptide sequences must match. Mappings shown in light grey are historical and correspond to inactive entries in the source databases.

Search Engine	Mascot
Score	63.89
Treshold	34.48
% Sequence Coverage	-

Quantitative Data

Source Name	Value
PRIDE TMT 126	6192618.999
PRIDE TMT 127	6454344.492
PRIDE TMT 128	7620348.236
PRIDE TMT 129	1.98224667E7
PRIDE TMT 130	1.659637806E7
PRIDE TMT 131	1.18205045E7



Additional

Source	Name	Value
User	MascotConfidenceLevel	95.0

Mapped Protein sequence

```
0001  MPSDLAKKKA AKKKEAAKAR QRPRKGHEEN GDVVTEPQVA EKNEANGRET TEVDLLTKEL EDFEMKKAAA RAVTGVLASH  0080
0081  PNSTDVHIIN LSLTFHGQEL LSDTKLELNS GRRYGLIGLN GIGKSMLLSA IGKREVPIPE HIDIYHLTRE MPPSDKTPLH  0160
0161  CVMEVDTERA MLEKEAERLA HEDAECEKLM ELYERLEELD ADKAEMRASR ILHGLGFTPA MQRKKLKDFS GGWRMRVALA  0240
0241  RALFIRPFML LLDEPTNHLD LDACVWLEEE LKTFKRILVL VSHSQDFLNG VCTNIIHMHN KKLKYYTGNY DQYVKTRLEL  0320
0321  EENQMKRFHW EQDQIAHMKN YIARFGHGSA KLARQAQSKE KTLQKMMASG LTERVVSDKT LSFYFPPCGK IPPPVIMVQN  0400
0401  VSFKYTKDGP CIYNNLEFGI DLDTRVALVG PNGAGKSTLL KLLTGELLPT DGMIRKHSHV KIGRYHQHLQ EQLDLDLSPL  0480
0481  EYMMKCYPEI KEKEEMRKII GRYGLTGKQQ VSPIRNLSDG QKCRVCLAWL AWQNPHMLFL DEPTNHLDIE TIDALADAIN  0560
0561  EFEGGMMLVS HDFRLIQQVA QEIWVCEKQT ITKWPGDILA YKEHLKSKLV DEEPQLTKRT HNVCTLTLAS LPRP
```

Do you want to know a bit more…?

D736–D742 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 11 November 2009*
doi:10.1093/nar/gkp964

The Proteomics Identifications database: 2010 update

Juan Antonio Vizcaíno[1], Richard Côté[1], Florian Reisinger[1], Harald Barsnes[2], Joseph M. Foster[1], Jonathan Rameseder[1,3], Henning Hermjakob[1] and Lennart Martens[1,*]

[1]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [2]Department of Informatics, University of Bergen, Norway and [3]Computational and Systems Biology Initiative, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received September 10, 2009; Revised October 6, 2009; Accepted October 13, 2009

ABSTRACT

The Proteomics Identifications database (PRIDE, http://www.ebi.ac.uk/pride) at the European Bioinformatics Institute has become one of the main repositories of mass spectrometry-derived proteomics data. For the last 2 years, PRIDE data holdings have grown substantially, comprising 60 different species, more than 2.5 million protein identifications, 11.5 million peptides and over 50 million spectra by September 2009. We here describe several new and improved features in PRIDE, including the revised submission process, which now includes direct submission of fragment ion annotations. Correspondingly, it is now possible to visualize spectrum fragmentation annotations on tandem mass spectra, a key feature for compliance with journal data submission requirements. We also describe recent developments in the PRIDE BioMart interface, which now allows integrative queries that can join PRIDE data to a growing number of biological resources such as Reactome, Ensembl, InterPro and UniProt. This ability to perform extremely powerful across-domain queries will certainly be a cornerstone of future bioinformatics analyses. Finally, we highlight the importance of data sharing in the proteomics field, and the corresponding integration of PRIDE with other databases in the ProteomExchange consortium.

INTRODUCTION

Mass spectrometry (MS) is currently the most commonly used technology for the identification and quantification of proteins. Like in any other 'omics' field, the amount of data generated by MS-based proteomics has increased exponentially in the last few years, which has prompted the development of several data repositories. The Proteomics Identifications database (PRIDE) (http://www.ebi.ac.uk/pride) was developed at the European Bioinformatics Institute (EBI), as a repository for the results of MS-based proteomics experiments, allowing data from a vast range of approaches, instruments and analysis platforms to be stored and disseminated in a common structured and queryable format. Originally established as a production service in 2005, PRIDE has previously been described (1–3) along with guidelines on using the database and its associated tools (4–6).

PRIDE does not stand alone in this field, however, as several other proteomics databases have been established over the past few years. GPMDB (7), PeptideAtlas (8) and Proteinpedia (9) are among the most important representatives of these (10). Additionally, the Tranche system (http://tranche.proteomecommons.org) provides a data transfer layer relying on peer-to-peer Internet protocol technology. Finally, the most recently launched proteomics repository is the NCBI Peptidome (11), a centralized, public proteomics data repository not dissimilar from PRIDE in its objectives. For an up-to-date review covering the capabilities of a comprehensive selection of proteomics MS repositories see Mead *et al.* (12).

PRIDE stores three different kinds of information: MS and MS/MS mass spectra as peak lists, the derived peptide and protein identifications (IDs) and any associated metadata. Indeed, one of the advantages PRIDE offers over other proteomics databases lies in the amount of structured metadata it contains, which is a key requirement to put the stored data in biological and/or technical context. Furthermore, together with the newly released NCBI Peptidome, the established PRIDE database constitutes an actual structured data repository, and does not assume any editorial control over submitted data.

Another important feature of PRIDE is that it allows data to remain private while anonymously sharing it with journal editors and reviewers through so-called 'reviewer log-in accounts'. As a result, PRIDE is now the recommended submission point for proteomics data for several journals such as *Nature Biotechnology* (13), *Nature*

*To whom correspondence should be addressed. Tel: +44 1223 492 610; Fax: +44 1223 494 484; Email: lennart.martens@ebi.ac.uk

CORRESPONDENCE

PRIDE Converter: making proteomics data-sharing easy

To the Editor:
Your editorial on 'Democratizing Proteomics Data'[1] correctly addressed the increasing importance of making proteomics data publicly available so that it can be audited, reanalyzed or reused. To make global data-sharing in the field work, however, it is important to minimize the burden of uploading data into publicly available databases, such as PRIDE[2]. To this end, we have written a freely available, open source tool called PRIDE Converter that makes it straightforward to submit proteomics data to PRIDE from most common data formats.

Public availability of data is the standard *modus operandi* for most of the life sciences, ranging from genome sequences, over microarray data, to protein information. Some of the best known examples in the field of proteomics include protein sequences in UniProt (http://www.uniprot.org/), protein structures in the Protein Databank (http://www.rcsb.org/) and protein modifications in UniMod and RESID (http://www.unimod.org/ and http://www.ebi.ac.uk/RESID). As highlighted in your 2007 editorial[1], making data publicly available in a standardized and structured way enables other researchers to access and reanalyze the data, and to use the collected results in novel ways.

Indeed, much of the progress over the past years in emerging fields, such as mass spectrometry (MS)-based proteomics, is directly related to the public availability of data obtained in earlier efforts[3], specifically the genome sequencing projects. Not surprisingly, the need for data-sharing in the field of proteomics itself was quickly pointed out[4]. Several proteomics MS data repositories have since been established, with GPMDB, PRIDE, PeptideAtlas and Proteinpedia among the most prominent[5]. With this infrastructure in place, journals have followed suit by starting to request deposition of MS-related data in these databases[1,6].

The PRIDE repository at the European Bioinformatics Institute (http://www.ebi.ac.uk/pride/) occupies a special place in the list of proteomics databases, in that it constitutes an actual data repository and

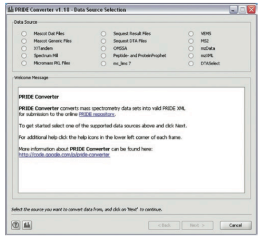


Figure 1 PRIDE Converter opening screen, showing the available input formats.

does not assume editorial control over submitted data[1]. Additionally, it provides a simple yet powerful infrastructure to support anonymous peer review of submitted data while maintaining the submission as private in the system[1]. The PRIDE database has so far accumulated data on more than 9,500 experiments, collectively containing more than 40 million mass spectra, identifying well over 1.4 million unique peptide sequences, which in turn infer more than 100,000 unique Ensembl proteins across all species.

Submitting an MS-based proteomics data set to a structured repository, such as PRIDE, has many advantages over alternative ways of making peptide and protein identifications publicly available, such as uploading raw data files on a web page[7] or submitting text or PDF files as supplementary information to a journal[2]. Furthermore, centralized repositories can also offer additional services and tools to the scientific community, based on uploaded data. PRIDE for instance includes tools for (i) visualizing protein coverage, peptide modifications and spectrum annotations, (ii) automatic mapping of protein accession numbers to identifiers from all other commonly used proteomics databases using the PICR service[8] and (iii) comprehensive protein list comparisons (through Venn diagrams)[9].

Submitting data to PRIDE could be challenging for some users, however. PRIDE relies on an XML-based data format for submissions, which is built around the Proteomics Standards Initiative mzData standard for mass spectrometry (http://www.ebi.ac.uk/pride/schemaXmlspyDocumentation.doi)[10]. And although the PRIDE XML format is well documented, converting proteomics data to PRIDE XML could present difficulties, especially for wet-lab scientists without a strong bioinformatics background or informatics support. To alleviate this problem, two tools for converting data into PRIDE XML have already been developed: the ProteomeHarvest PRIDE Submission Spreadsheet, which is Microsoft Excel-based (http://www.ebi.ac.uk/pride/proteomeharvest/), and the PRIDE Wizard for Mascot result files (http://www.mcisb.

Proteomics 2009, *9*, 1–8 DOI 10.1002/pmic.200900402 1

STANDARDISATION & GUIDELINES

A guide to the Proteomics Identifications Database proteomics data repository

Juan Antonio Vizcaíno[1], Richard Côte[1], Florian Reisinger[1], Joseph M. Foster[1], Michael Mueller[1], Jonathan Rameseder[1,2], Henning Hermjakob[1] and Lennart Martens[1]

[1] EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[2] Computational and Systems Biology Initiative, Massachusetts Institute of Technology, Cambridge, MA, USA

The Proteomics Identifications Database (PRIDE, www.ebi.ac.uk/pride) is one of the main repositories of MS derived proteomics data. Here, we point out the main functionalities of PRIDE both as a submission repository and as a source for proteomics data. We describe the main features for data retrieval and visualization available through the PRIDE web and BioMart interfaces. We also highlight the mechanism by which tailored queries in the BioMart can join PRIDE to other resources such as Reactome, Ensembl or UniProt to execute extremely powerful across-domain queries. We then present the latest improvements in the PRIDE submission process, using the new easy-to-use, platform-independent graphical user interface submission tool PRIDE Converter. Finally, we speak about future plans and the role of PRIDE in the ProteomExchange consortium.

Received: June 9, 2009
Revised: June 24, 2009
Accepted: June 25, 2009

Keywords:
Bioinformatics / Data repository / Mass spectrometry

1 Introduction

Bioinformatics tools and data repositories provide one of the main pillars of biology in the 21st century. Indeed, public availability of biological data *via* the Internet has changed the way biologists plan, execute and interpret their studies. Some of the best known protein-related resources include UniProt [1] for protein sequences and annotation, the Protein Databank [2] and other members of the wwPDB consortium [3] for protein structures, Intact [4] and other components of the IMEx consortium [5] for protein interactions, InterPro [6] for protein domains, and UniMod [7] and RESID [8] for protein modifications.

Like in any other "omics" field, the amount of data generated by MS based proteomics has increased exponentially in the last few years, which prompted the development of several data repositories. At the same time, proteome informatics efforts have driven the development of universally adopted and stable data formats under the auspices of the HUPO Proteomics Standards Initiative (HUPO-PSI, www.psidev.info), and have led to powerful data analysis strategies [9, 10]. Taken together, these advances have allowed the centralized aggregation of proteomics data and its reanalysis or meta-analysis, ultimately turning proteomics into a much more robust discipline in the life sciences.

Several proteomics MS data repositories have been established so far, with GPMDB [11], Proteomics Identifications Database (PRIDE) [12], PeptideAtlas [13] and Proteinpedia [14] among the most prominent ones at present [15, 16]. Additionally, the NCBI recently launched their Peptidome (http://www.ncbi.nlm.nih.gov/projects/peptidome) system as a centralized, public proteomics repository not dissimilar from PRIDE. The Tranche (http://tranche.proteomecommons.org) system is used in the field as well, and essentially presents a data transfer layer relying on peer-to-peer Internet protocol technology. Apart from these large-scale efforts, there are also smaller, more

Correspondence: Dr. Lennart Martens, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
E-mail: lennart.martens@ebi.ac.uk
Fax: +44-1223-494-484

Abbreviations: EBI, European Bioinformatics Institute; **OLS,** Ontology Lookup Service; **PICR,** Protein Identifier Cross-Referencing; **PRIDE,** Proteomics Identifications Database; **PSI,** Proteomics Standards Initiative

Juan A. Vizcaíno
juan@ebi.ac.uk

BSPR/EBI Educational Workshop
Hinxton, 16 July 2010

EMBL-EBI

The PRIDE Team

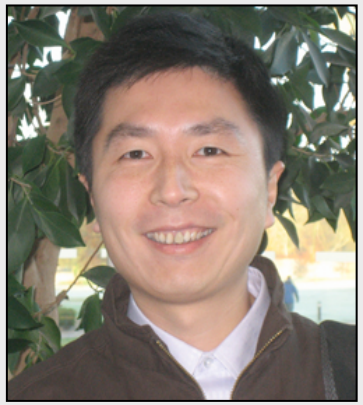
Attila Csordas

Daniel Ríos

Richard Côté

Florian Reisinger

Henning Hermjakob



Rui Wang

Joe Foster
(Ph.D. student)

Andreas Schonegger
(Trainee)

Links, collaborations and funding

http://www.psidev.info

http://www.ebi.ac.uk/ols

http://www.ebi.ac.uk/pride

http://www.ebi.ac.uk/tools/picr

http://www.ebi.ac.uk/pride/dod

juan@ebi.ac.uk

pride-support@ebi.ac.uk

PRIDE collaborators

Yonsei Proteome Research Center

UNIVERSITEIT GENT

MANCHESTER 1824

Funding

wellcome trust

EMBL-EBI

bbsrc biotechnology and biological sciences research council

Thank you!

Questions?