# Proteomics Tools & Data Archive Construction

Ronald Beavis

What protein sequences do we use to maintain references to external database resources?

Proteomics initially was forced to use "phenomenological" lists of proteins.

For example: NCBI's non-redundant sequence collection
non-redundant = exactly the same sequence was not repeated

This type of list was necessary, because there were no sequenced genomes in existence for model organisms and the protein sequences that did exist contained significant errors.

$\mathbf{P} = \{P_j\}$ - a set of protein sequences

What set of proteins should be used?

Proteomics can now use translated gene models for most major experimental species.

1. ENSEMBL & TIGR collections (genomic);
2. UNIGENE (EST contigs, transcriptomic);
3. Boutique organism-based sites (SGD, FlyBase)

These sequence collections are non-redundant in the biological sense, but they do pose some problems.

Phenomenological sequence collections attempt to find every possible gene product sequence. Gene model predictions must be manipulated to find all possible gene products.

Enter the genome

$\mathbf{P} = \{\mathbf{T}_j \otimes \mathbf{R}_j \otimes \mathbf{D}_j \otimes g_i\}$ - derived from genes ($g_i$)

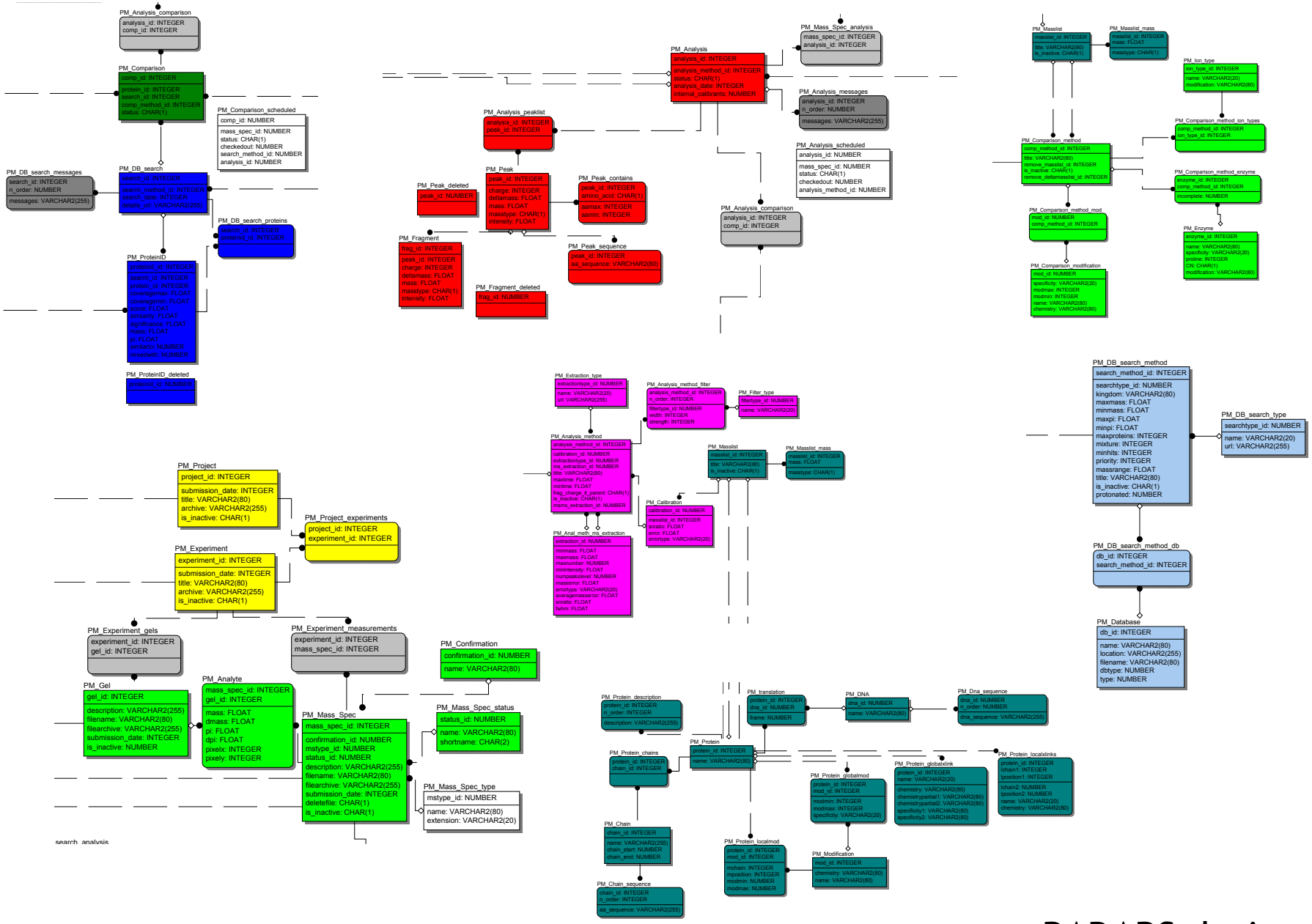
$\mathbf{D}_j$ – DNA operations (deletion, insertion, mutation)

$\mathbf{R}_j$ – RNA operations (splicing)

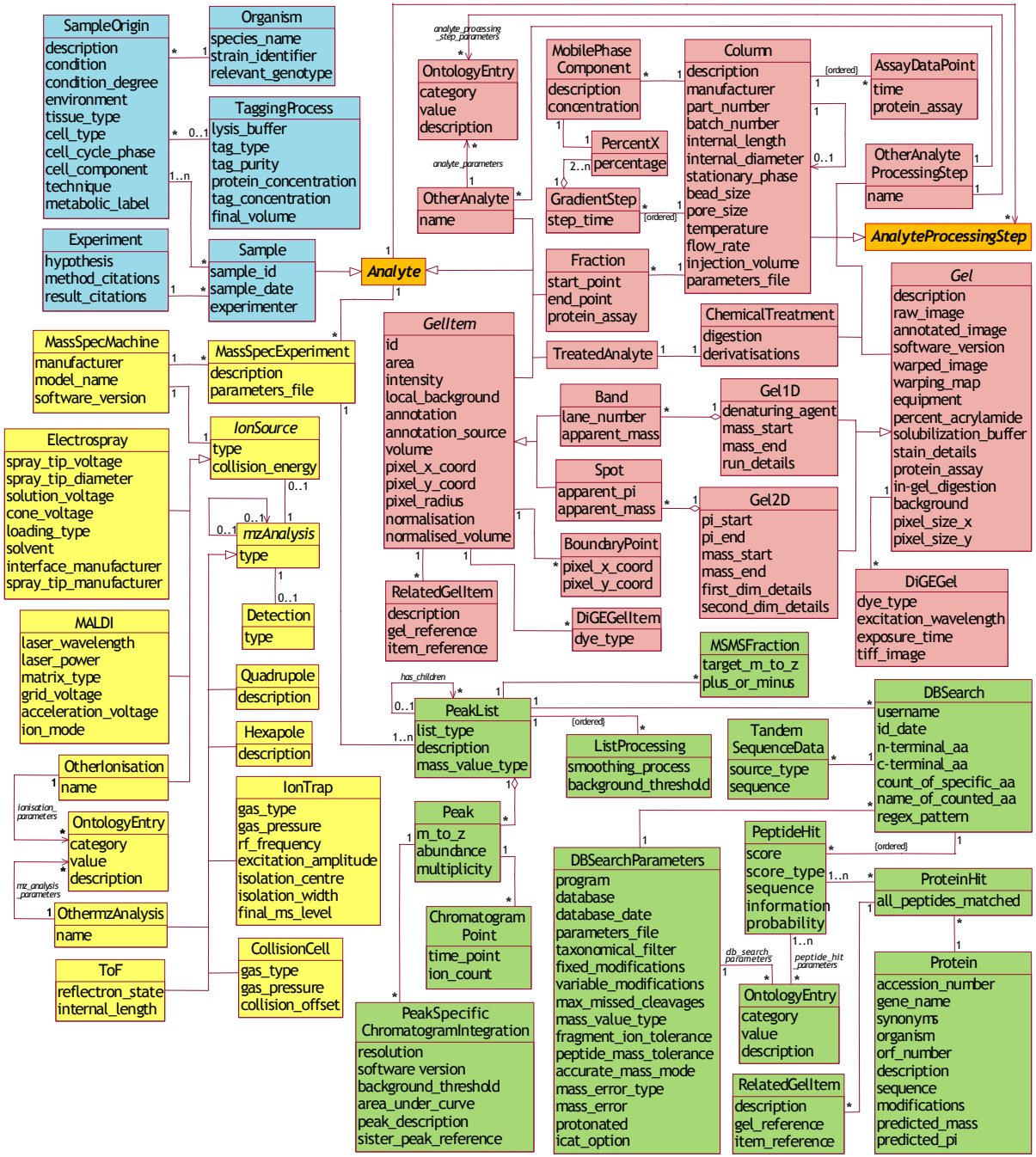
$\mathbf{T}_j$ – protein operations (cleavage, PTM)

Operations that lead to proteins

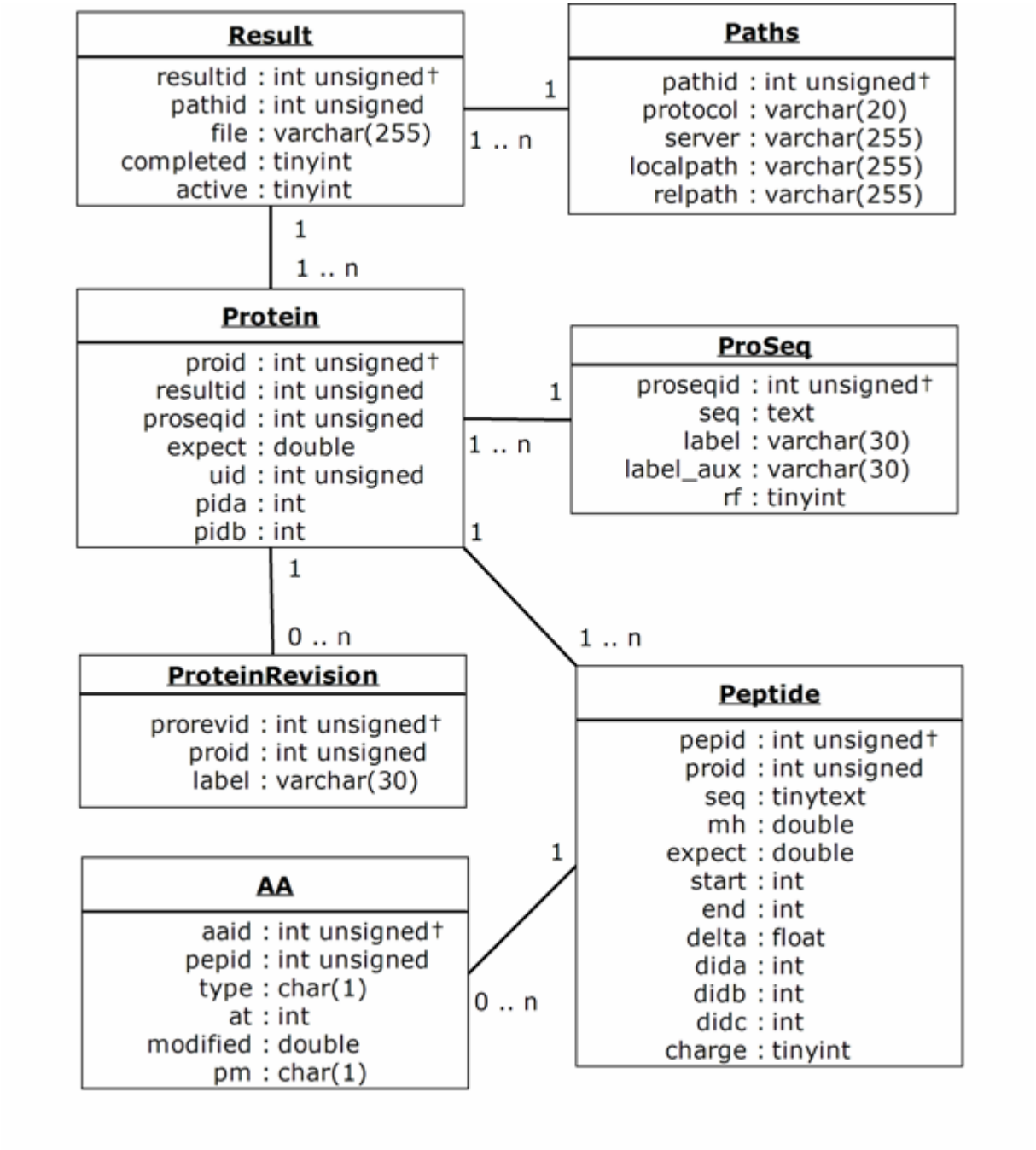
What information do we store in a data archive and how should it be referenced?



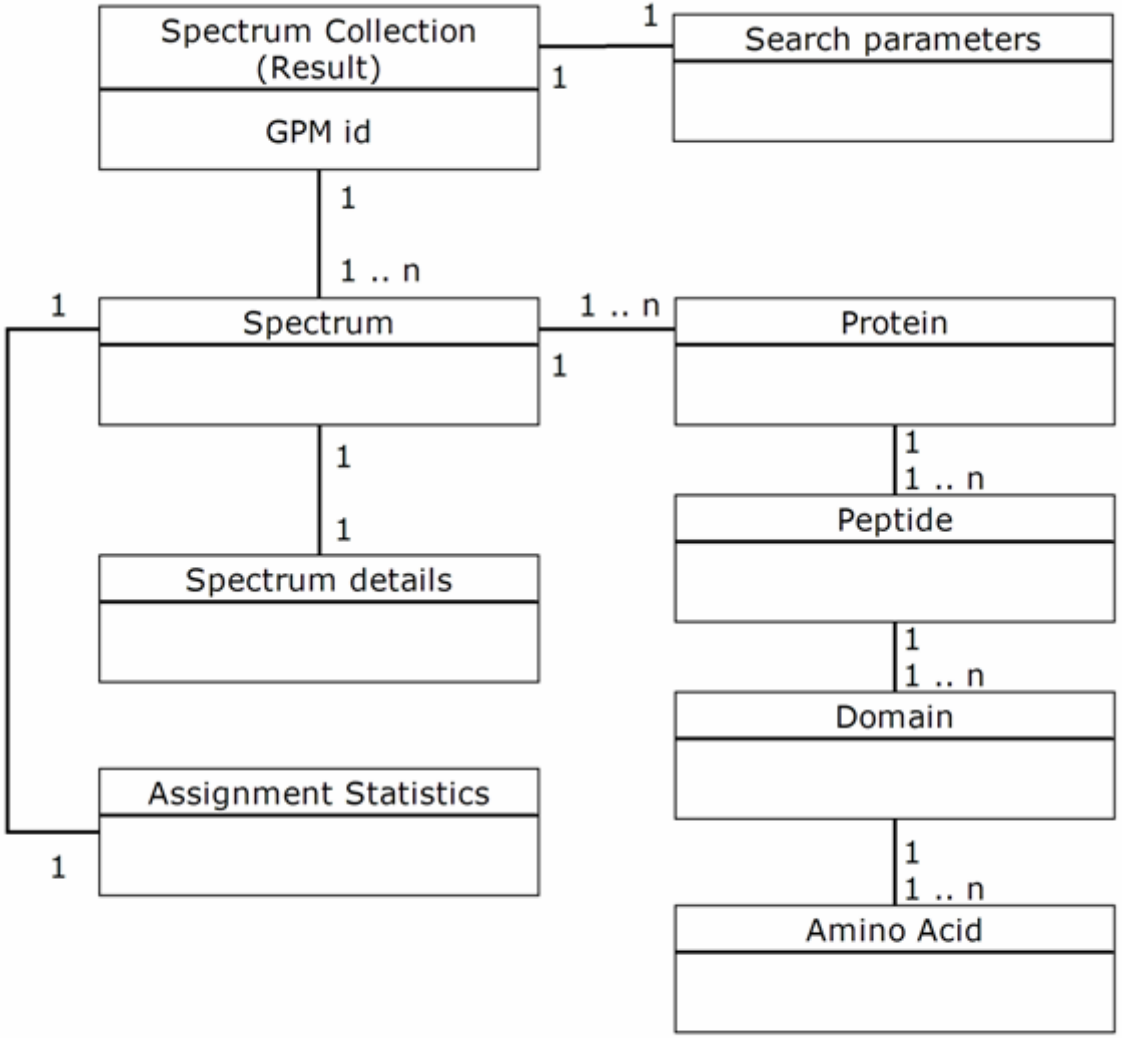

# RADARS design



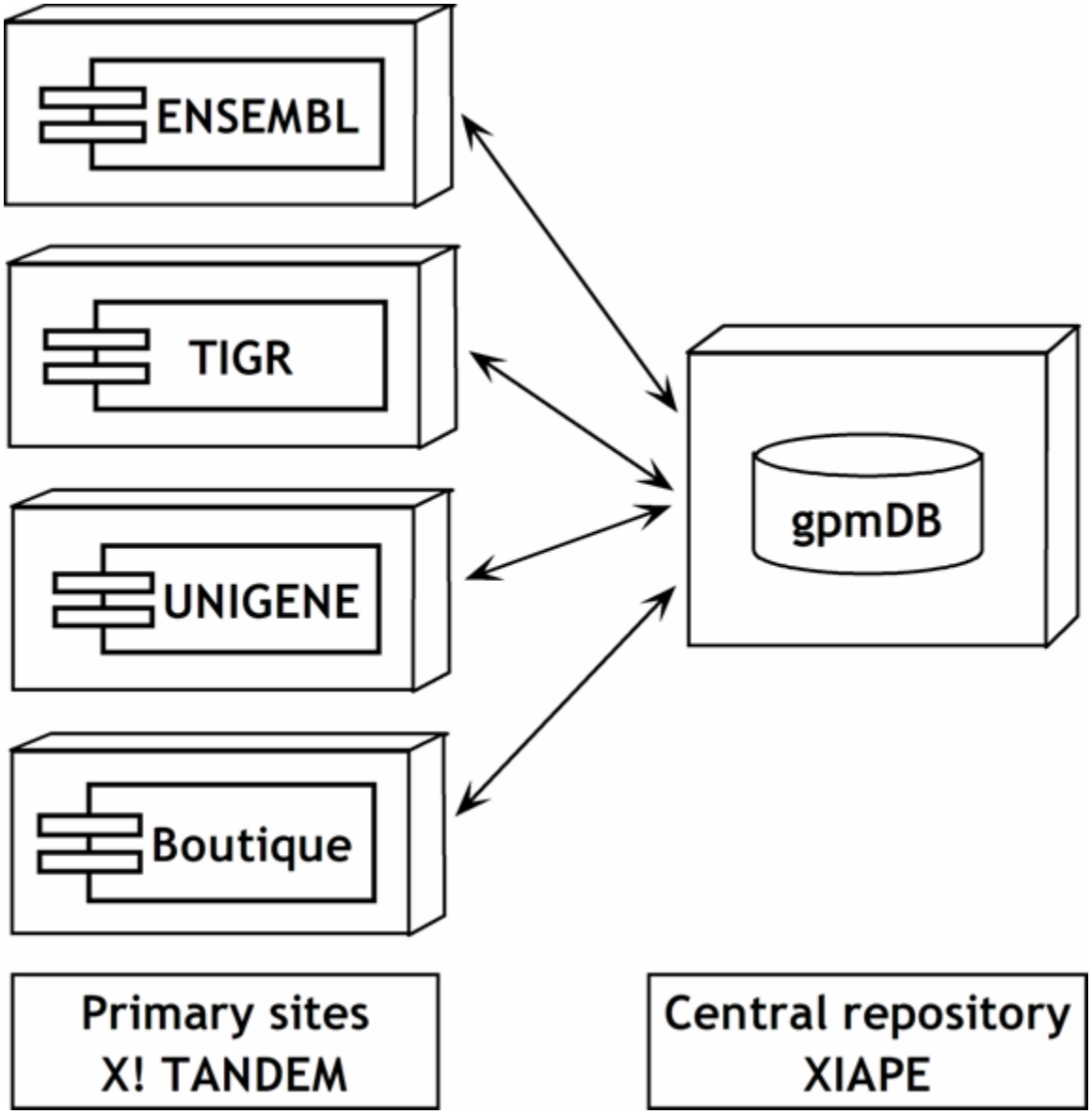

# MIAPE design




GPMDB Database design




GPMDB XML Archive Design

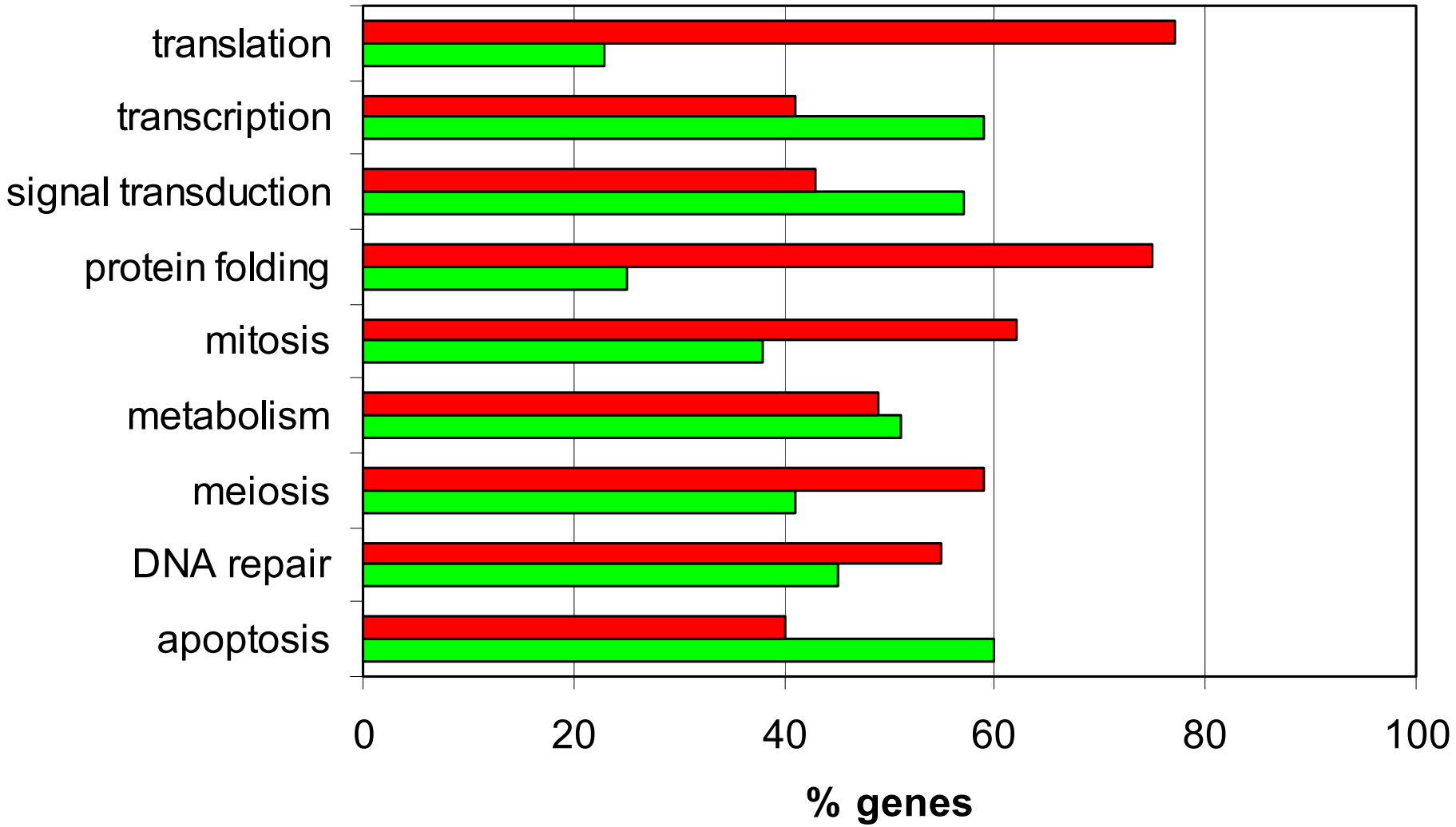



Deployment design

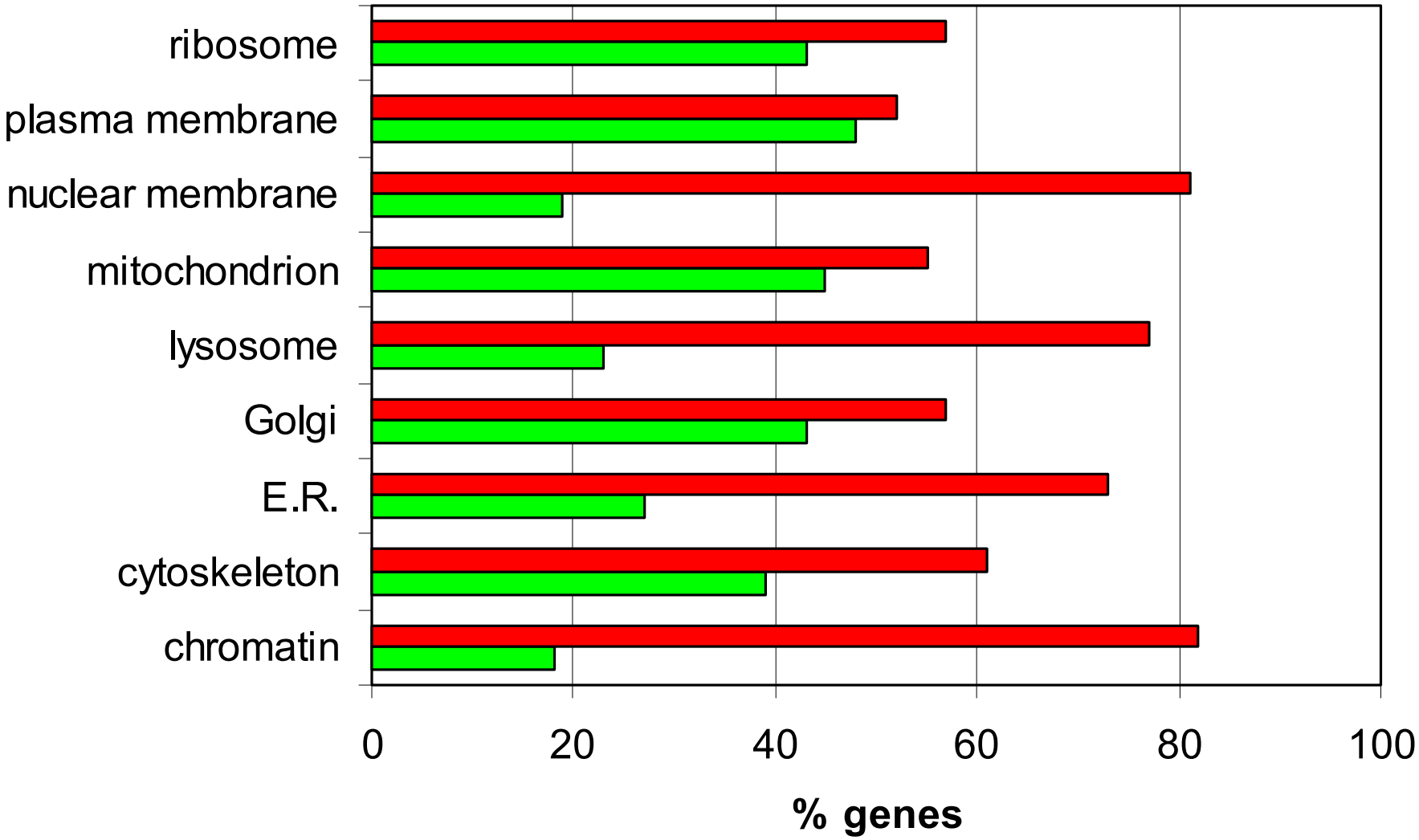
# How do you use a data archive as a research tool?

| | | |
|---|---|---|
| models | = | 12217 |
| proteins | = | 697807 |
| protein redundancy | = | 8.2× |
| peptides | = | 4,463,690 |
| peptide redundancy | = | 20.2× |

GPMDB Archive size



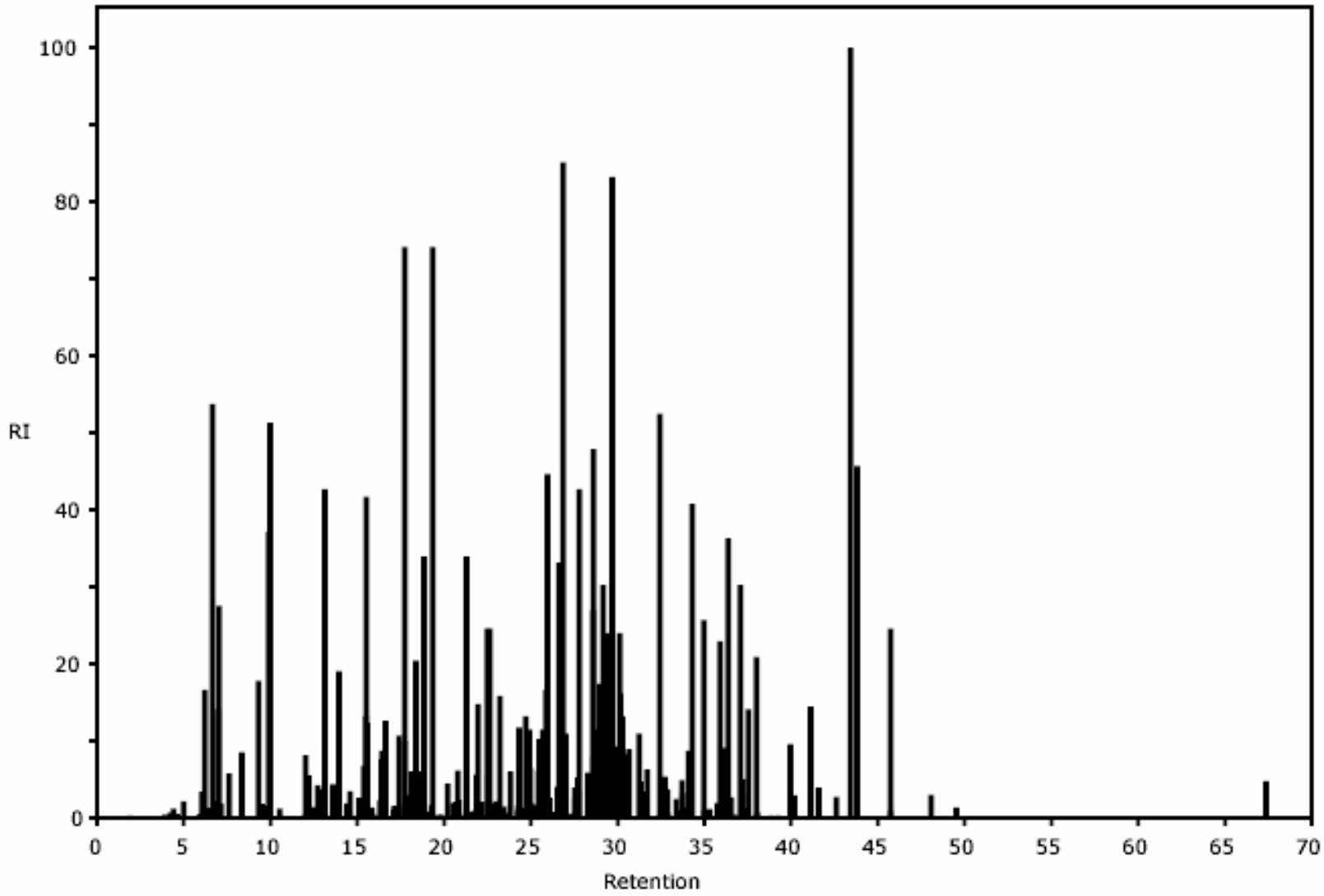

Functional ontology snapshot of GPMDB (Dec. 20, 2004)



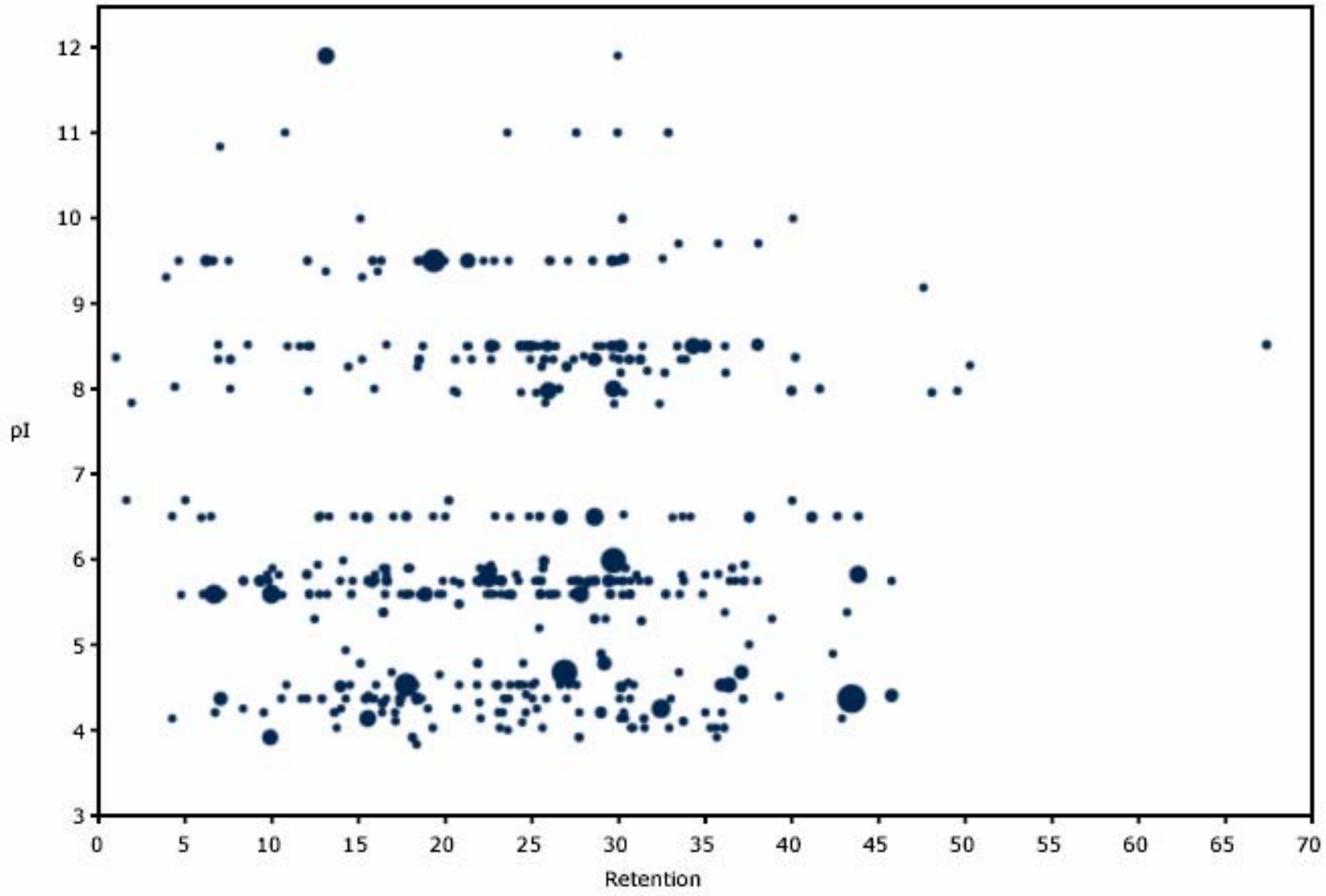

Cell component ontology snapshot of GPMDB (Dec. 20, 2004)

# Data representation:

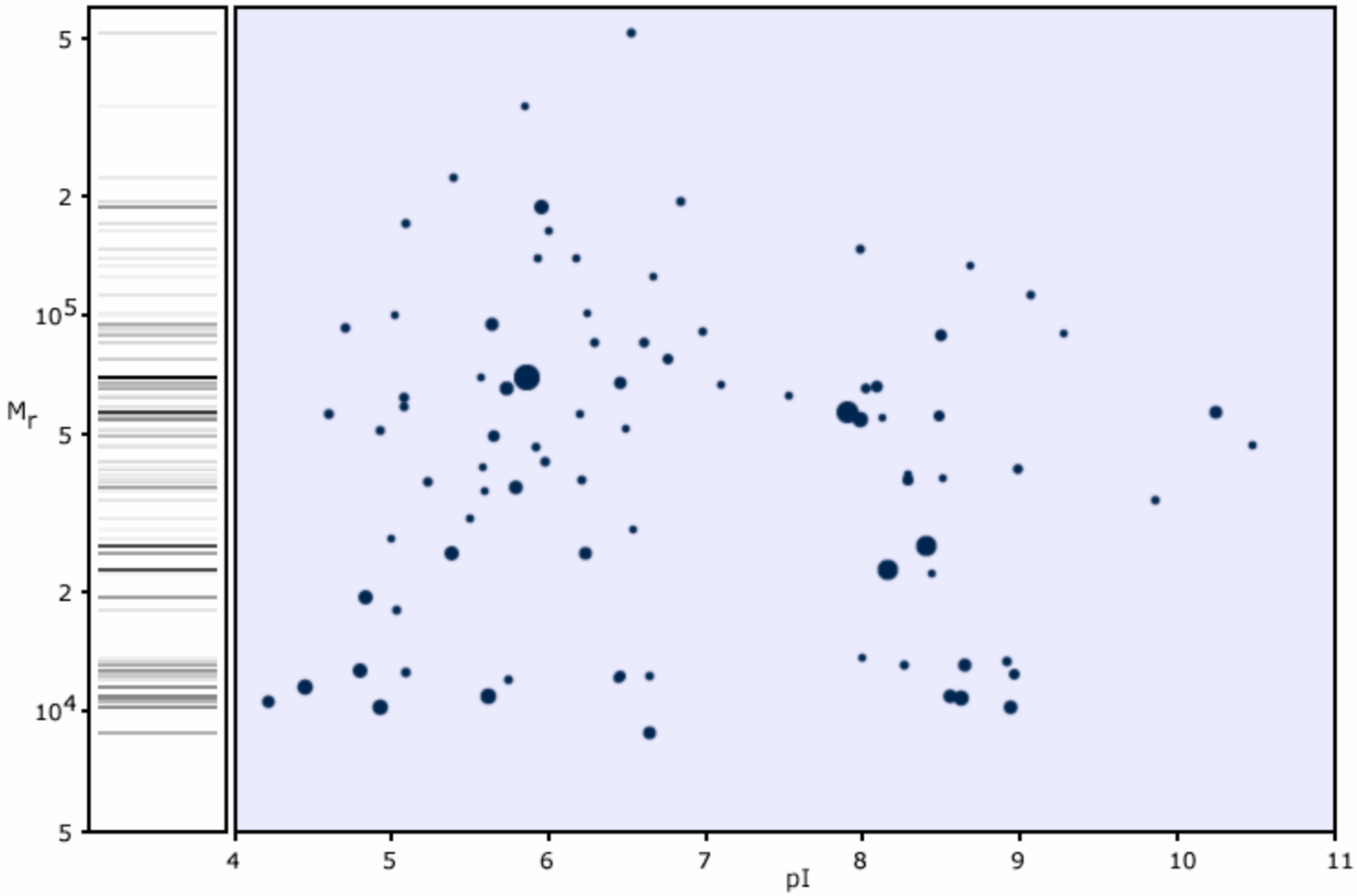
What types of diagrams are useful in evaluating and validating proteomics results.




Simulated HPLC




Simulated pI vs retention time




Simulated 1D & 2D gels

**ENSP00000256119:**
**log(e) = -1.2**

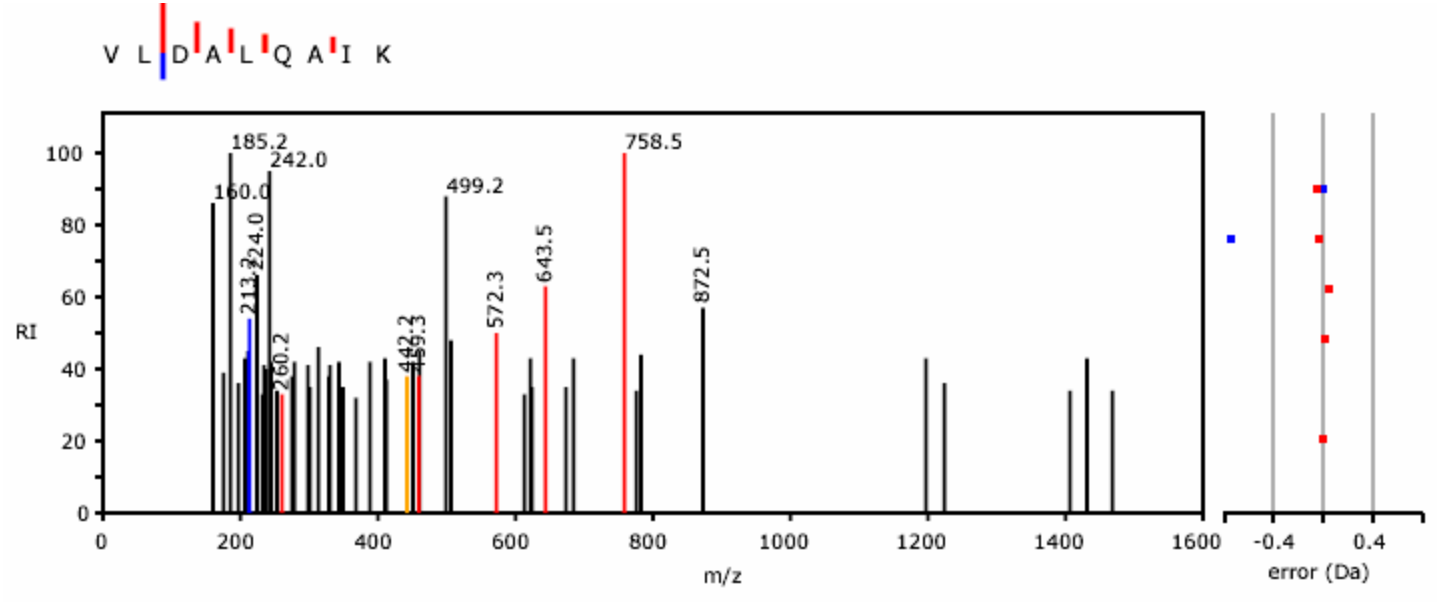
Carbonic anhydrase I (EC 4.2.1.1) (Carbonate dehydratase I) (CA-I) (Carbonic anhydrase B).
[Source:Uniprot/SWISSPROT;Acc:P00915]

Annotated domains:

IPR001148 Carbonic anhydrase, eukaryotic

```
1   maspdwgyddkngpeqwsklypiangnnQspvdiktsetkhdtslkpisvsynpatakei   60
    MASPDWGYDDKNGPEQWSKLYPIANGNNQSPVDIKTSETKHDTSLKPISVSYNPATAKEI

61  invghsfhvnfedndnrsvlkggpfsdsyrlfqFhfhwgstnehgsehtvdgvkysaelh  120
    INVGHSFHVNFEDNDNRSVLKGGPFSDSYRLFQFHFHWGSTNEHGSEHTVDGVKYSAELH

121 vahwnsakysslaeaaskadglAvigvlmkvgeanpklqkvldalqaiktkgkrapftnf  180
    VAHWNSAKYSSLAEAASKADGLAVIGVLMKVGEANPKLQKVLDALQAIKTKGKRAPFTNF

181 dpstllpssldfwtypgslthpplyesvtwiickesisvsseqlaqfrsllsnvegdnav  240
    DPSTLLPSSLDFWTYPGSLTHPPLYESVTWIICKESISVSSEQLAQFRSLLSNVEGDNAV

241 pmqhnnrptqplkgrtvrasf                                         261
    PMQHNNRPTQPLKGRTVRASF
```

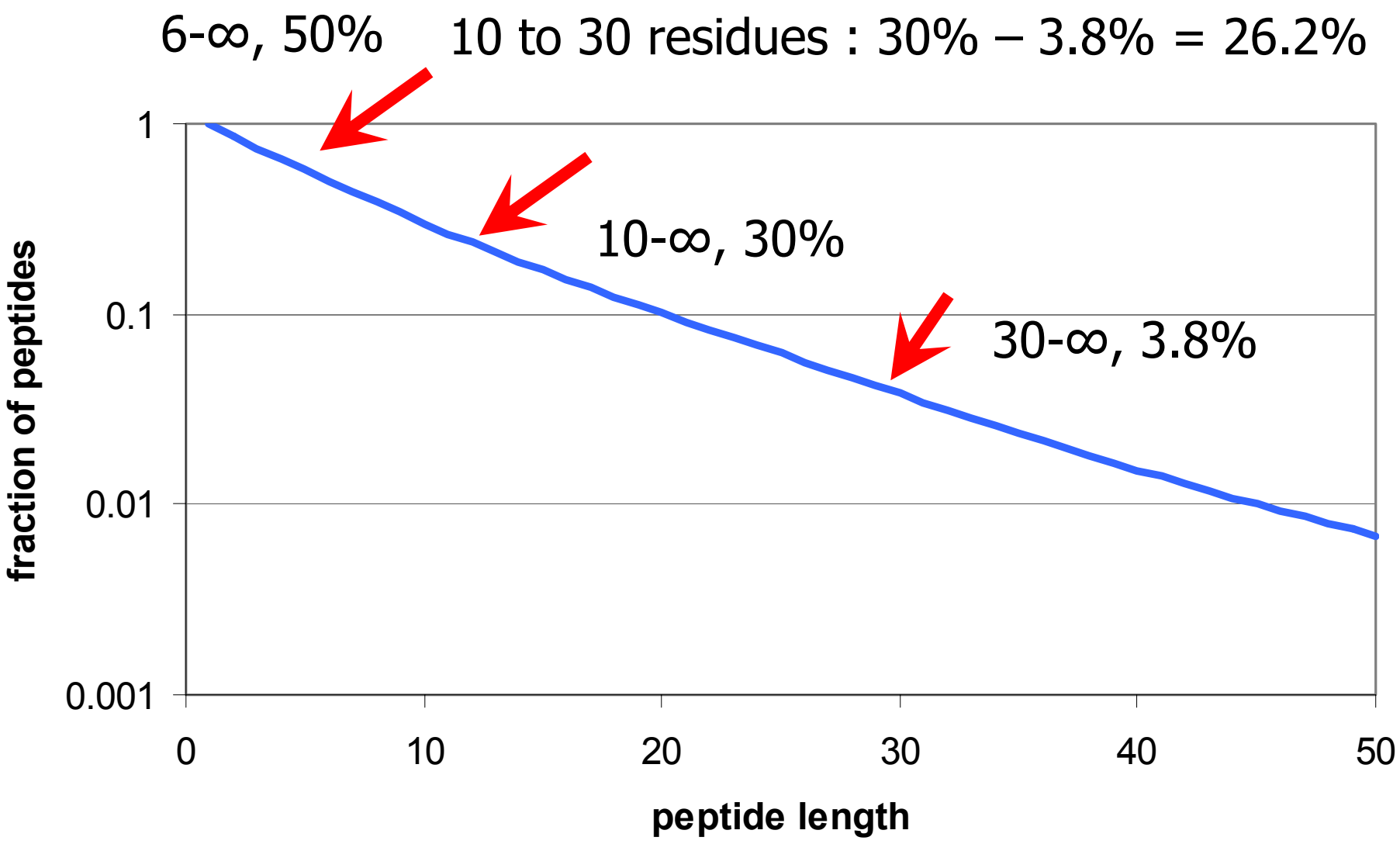
Validating data



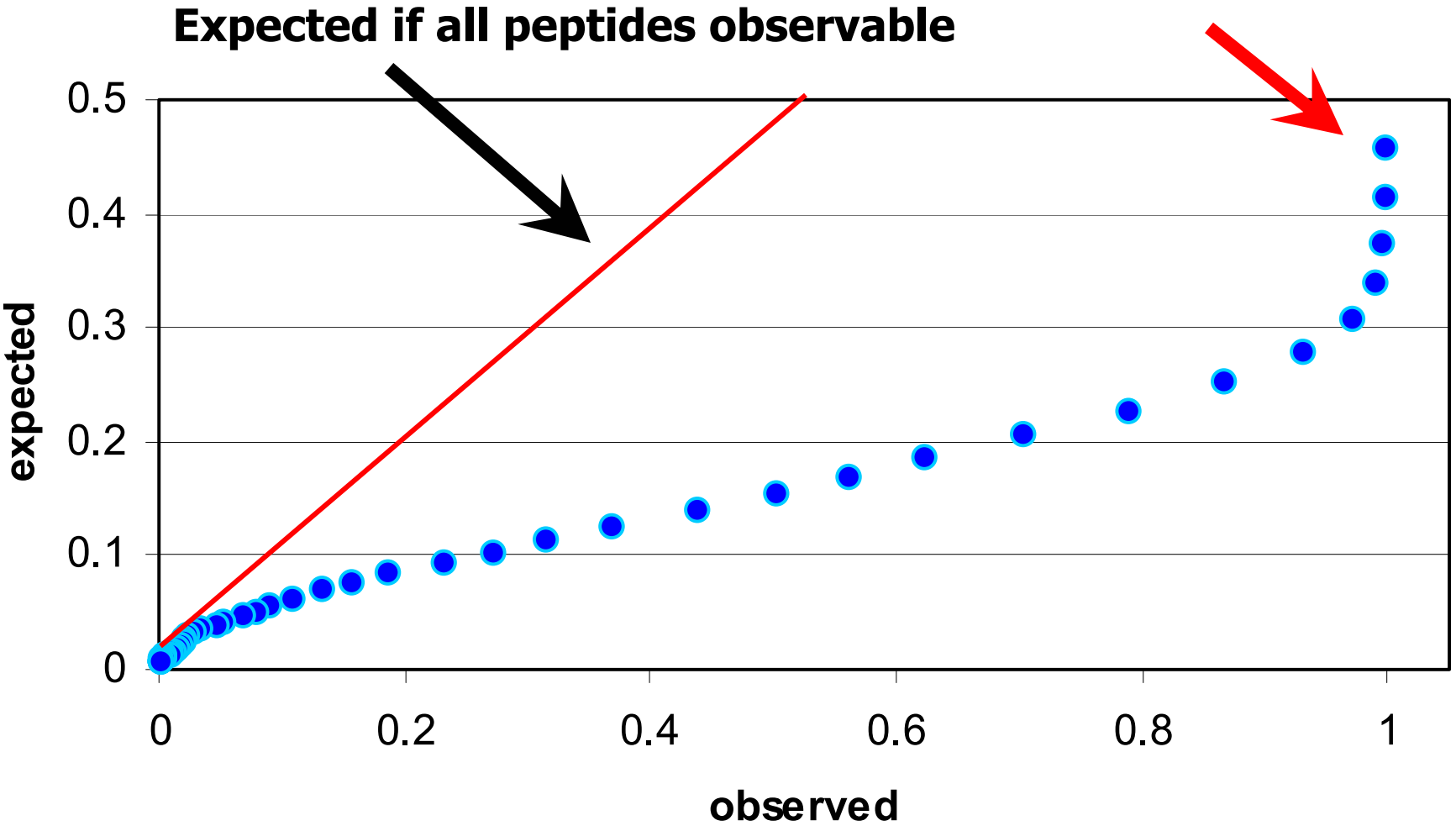

Comparing spectra


Comparing coverage

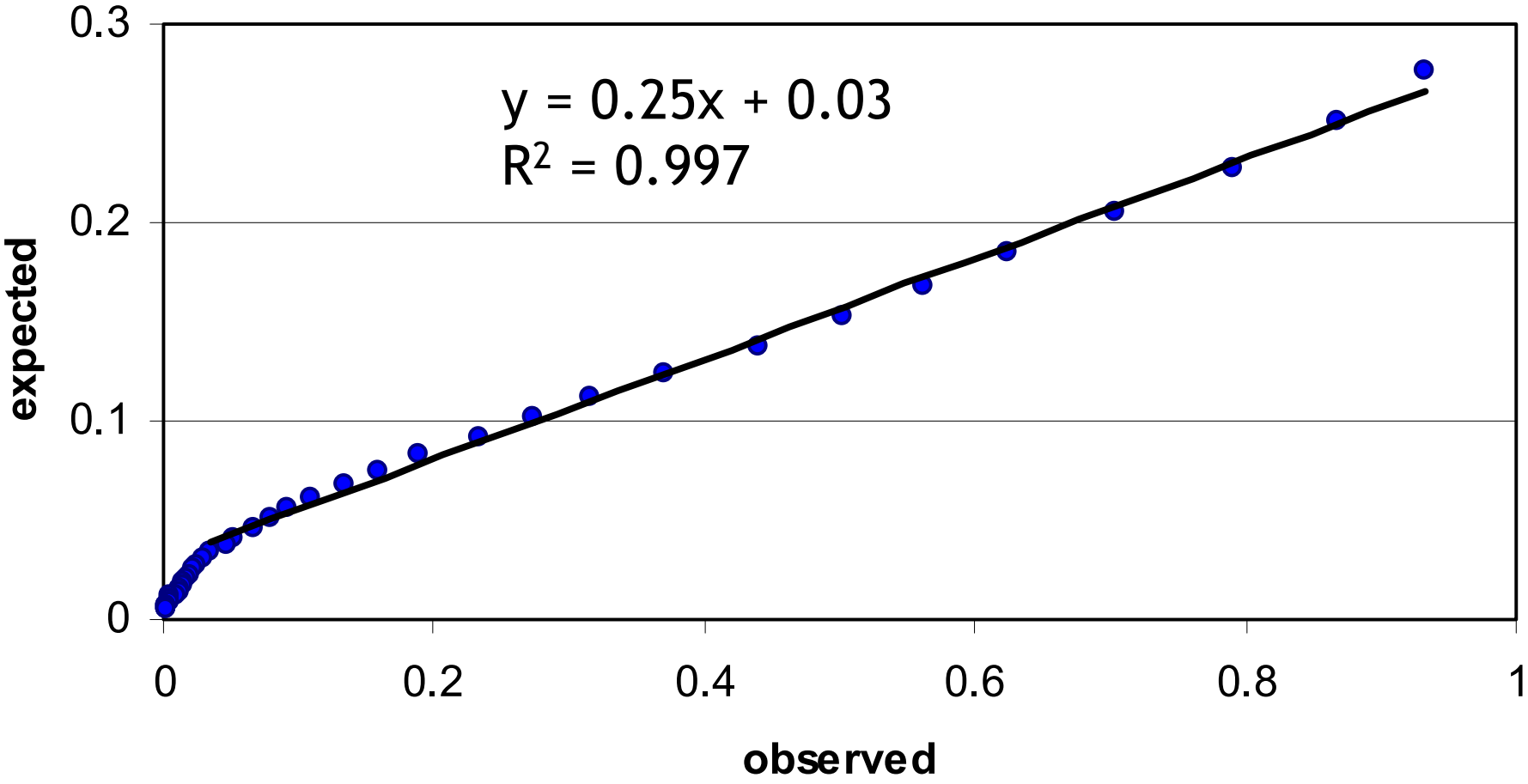
Can you use an archive to speed up identifications?



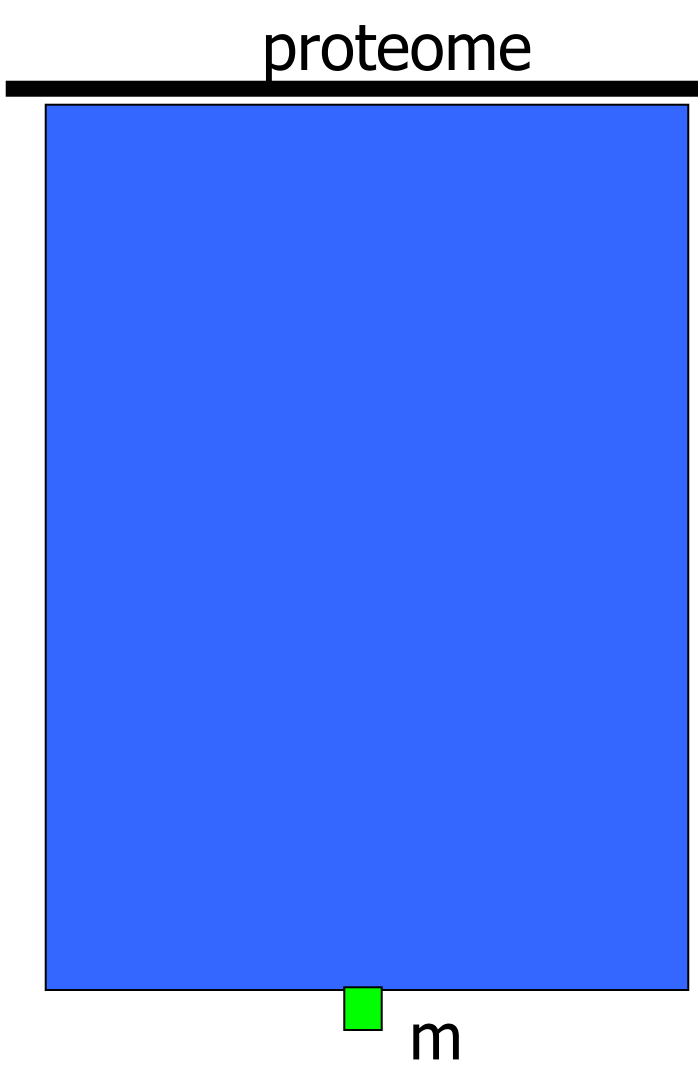

Proteome "coverage": counting tryptic peptides




Comparison of expected vs. observed peptides




Comparison of expected vs. observed peptides




model sequences

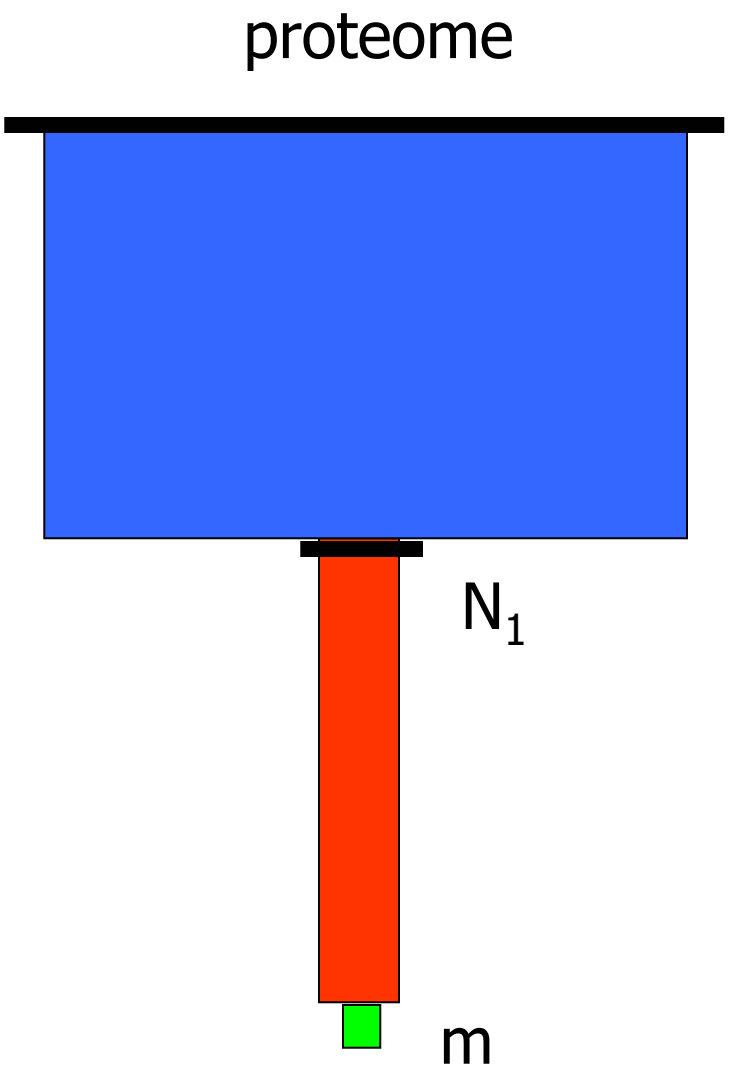
All analysis done on all sequences, leading to sequence models

Current strategy employed for analyzing spectra v. sequences

For each identifiable protein in an original protein mixture, there will exist at least one detectable tryptic peptide with $n \leq n_{R1}$.

*n is the number of missed tryptic cleavages*
*$n_{R1}$ is a constant (e.g. $n_{R1} = 1$)*

An approximation for a collection of ms/ms spectra




1st round: fast
creates initial models

---

$2^{nd}$ round
$3^{rd}$ round
…
$n^{th}$ round
refines the models

Refinement strategy for analyzing spectra v. sequences

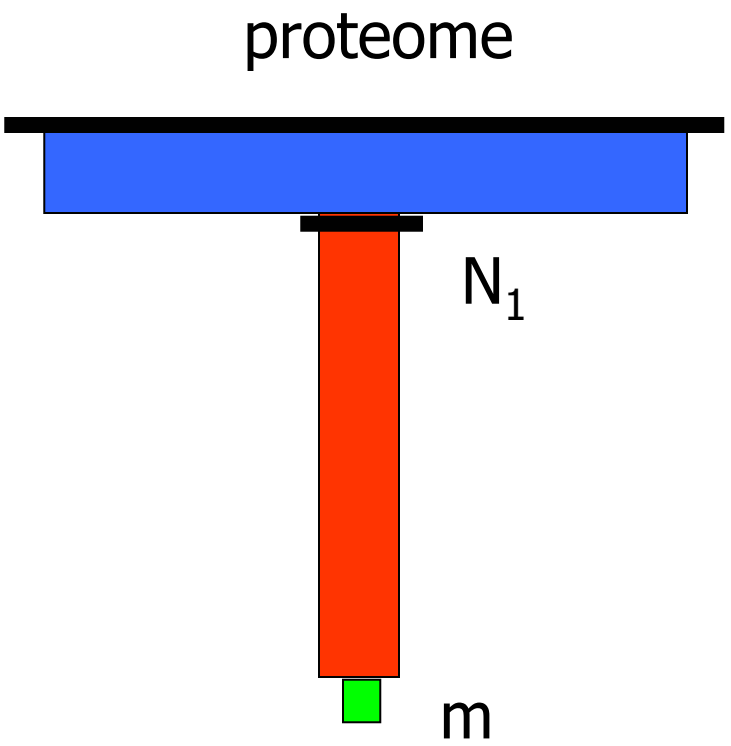
1. Incremental rebuild of GPMDB with all data collected the day before.

2. Find all observed peptides for a particular species, with $e < e_{max}$

3. Store peptide sequences, associated with the protein sequence accession number.

4. Update search sequence list on search machines.

Improved sequence assignment strategy




1st round: very fast
Finds accession numbers

---

$2^{nd}$ - $n^{th}$ round
Gets full sequences and fully analyzes them

Refinement strategy for analyzing spectra v. sequences

| **Parameter** : | **Value** |
|---|---|
| total spectra assigned: | 725 |
| total spectra used: | 6863 |
| total unique assigned: | 381 |
| initial modeling total (sec): | 3.016 |
| initial modeling/spectrum (sec): | 0.00044 |
| refinement/spectrum (sec): | 0.00012 |

Search results, using Dell 470 dual processor workstation

Rob Craig
John Cortens

All of the groups that have submitted their data and who have set up their own installations of these systems.