SpectraST:
A Spectral Library Searching Tool for Proteomics

Henry Lam

Day 3

October 18, 2006

INSTITUTE FOR Systems Biology

Why Spectral Searching?

- Traditional sequence (database) searching is
  - Very computationally intensive and costly
  - Error prone
  - Unable to capitalize on past data
  - Good for purely discovery-oriented experiments
- Newer approaches to proteomics are often more *targeted*
  - Know what you are looking for
  - More interest in probing/quantifying/understanding proteome segments that have already been mapped out
  - Repeated sampling of same proteome segments

Spectral Library Searching

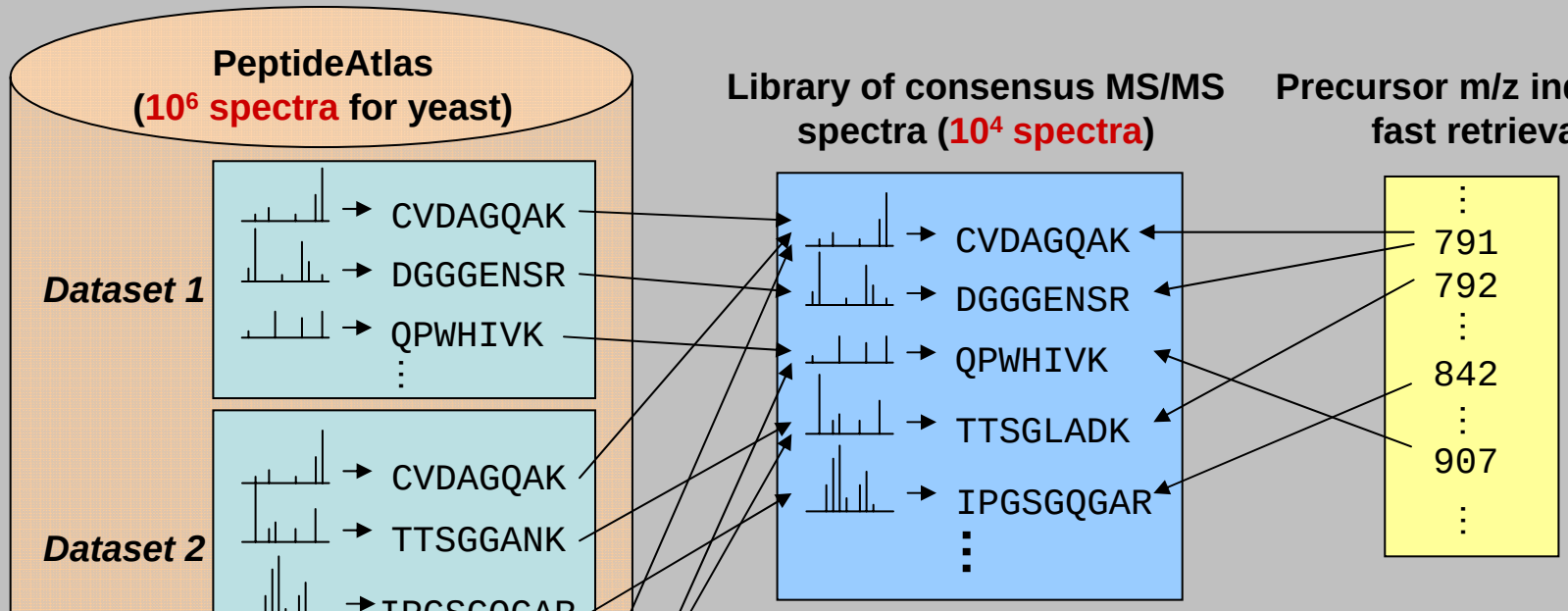
- Identifying an unknown peptide MS/MS (CID) spectrum by matching it against a library of known peptide MS/MS spectra
- Premise: *One-to-one correspondence* between peptide ion (sequence + charge + modifications) and its characteristic MS/MS “fingerprint”
- General and widely practiced method for small molecules
  - Problem with proteomics until recently: lack of good spectral libraries*

* Yates, J. R., et al., *Anal. Chem.* 1998, **70**, 3557-3565.

Spectral Libraries

- Impractical to construct spectral libraries using purified peptides; instead, *peptide-spectrum correspondence established by sequence searching* on complex samples
- Recent developments enabling the construction of peptide spectral libraries
  - Explosion of shotgun proteomics data
  - Emergence of public data repositories
  - Standardization of data formats
- *Consensus spectral libraries*: Multiple observations of the same peptide ion aggregated to form “consensus spectra”
  - Decrease library size
  - Give higher confidence in peptide-spectrum correspondence
  - Average out experiment-to-experiment variations
  - Reduce noise and other spurious peaks (e.g. from impurities)

Spectral Library Creation



1. Search datasets by ***4 different sequence search engines*** (SEQUEST, Mascot, X!Tandem, OMSSA)
2. Group ***replicate spectra*** identified to same peptide ion with high confidence
3. Combine replicates to create ***consensus spectra***
4. Apply ***quality filters*** to consensus spectra
5. Build searchable libraries with indexes

* Details described in NIST library documentation

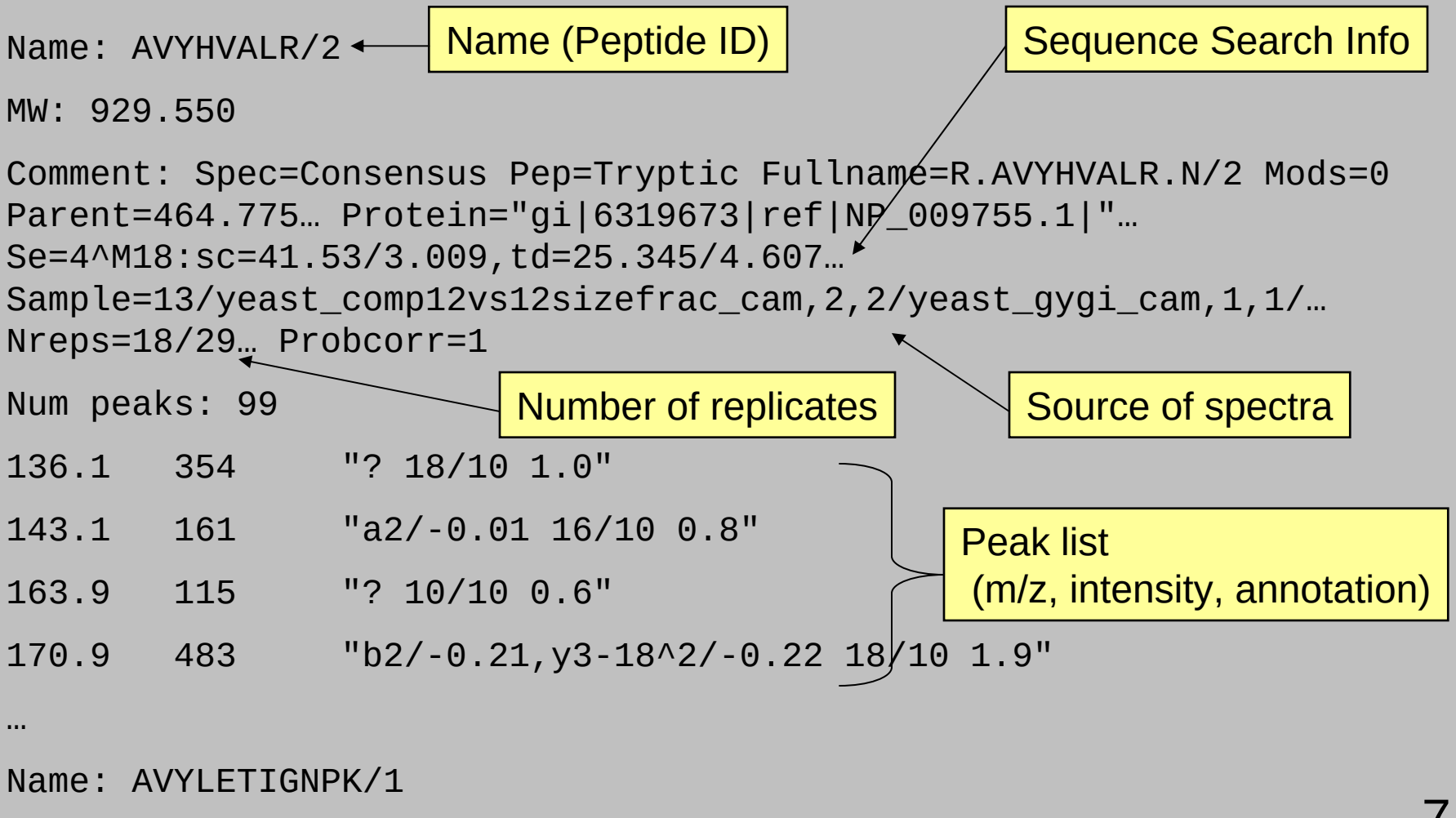
NIST Libraries

- NIST consensus spectral libraries (as of Sept 2006)
  - Available on PeptideAtlas and ProteomeCommons

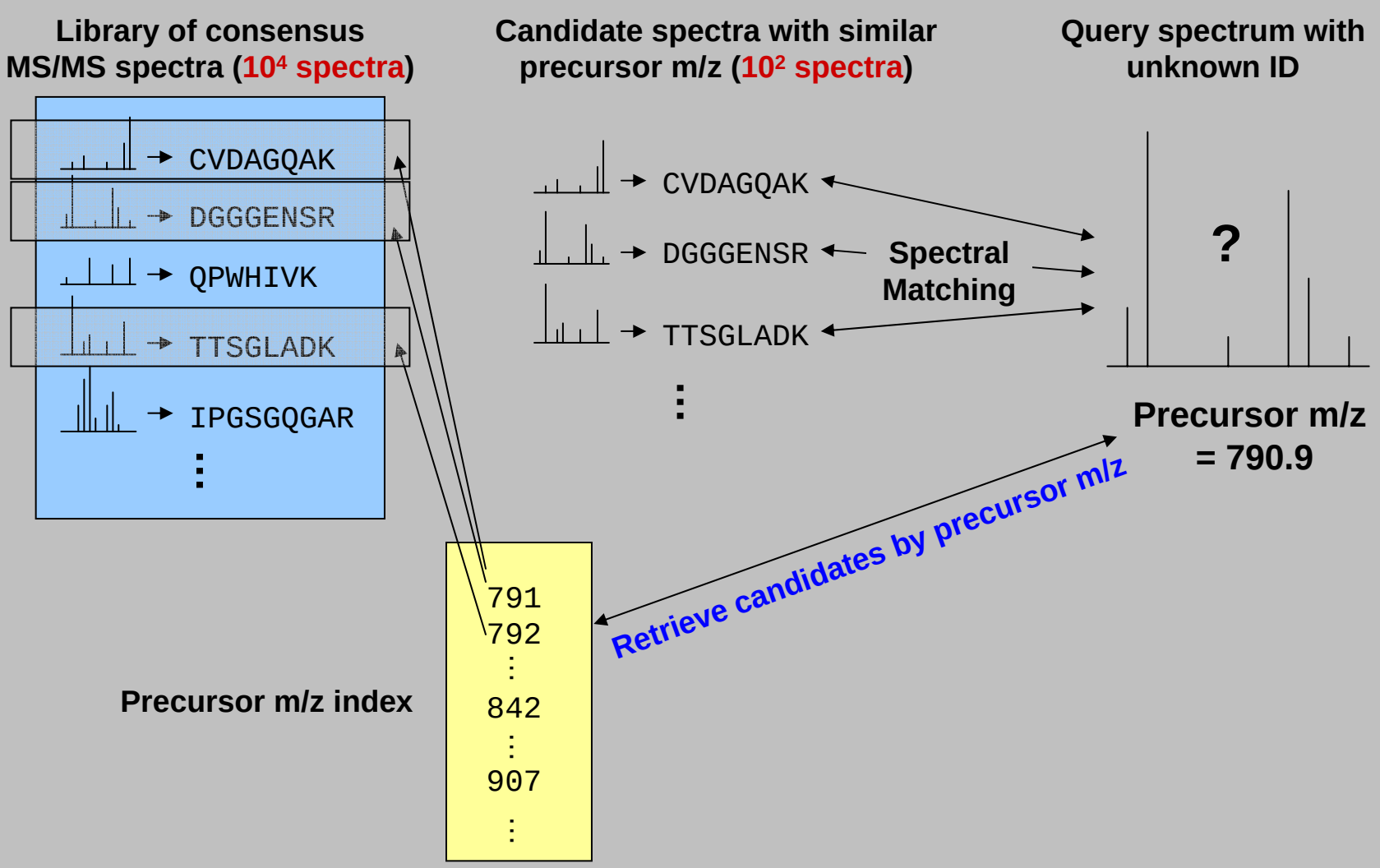
Organism	# datasets used	# consensus spectra
Yeast	43	35,135
Human	77	45,377
D. radiodurans	1	8,273
M. smegmatis	1	3,569
19 Standard proteins	28	3,938

NIST Libraries

- In .msp format



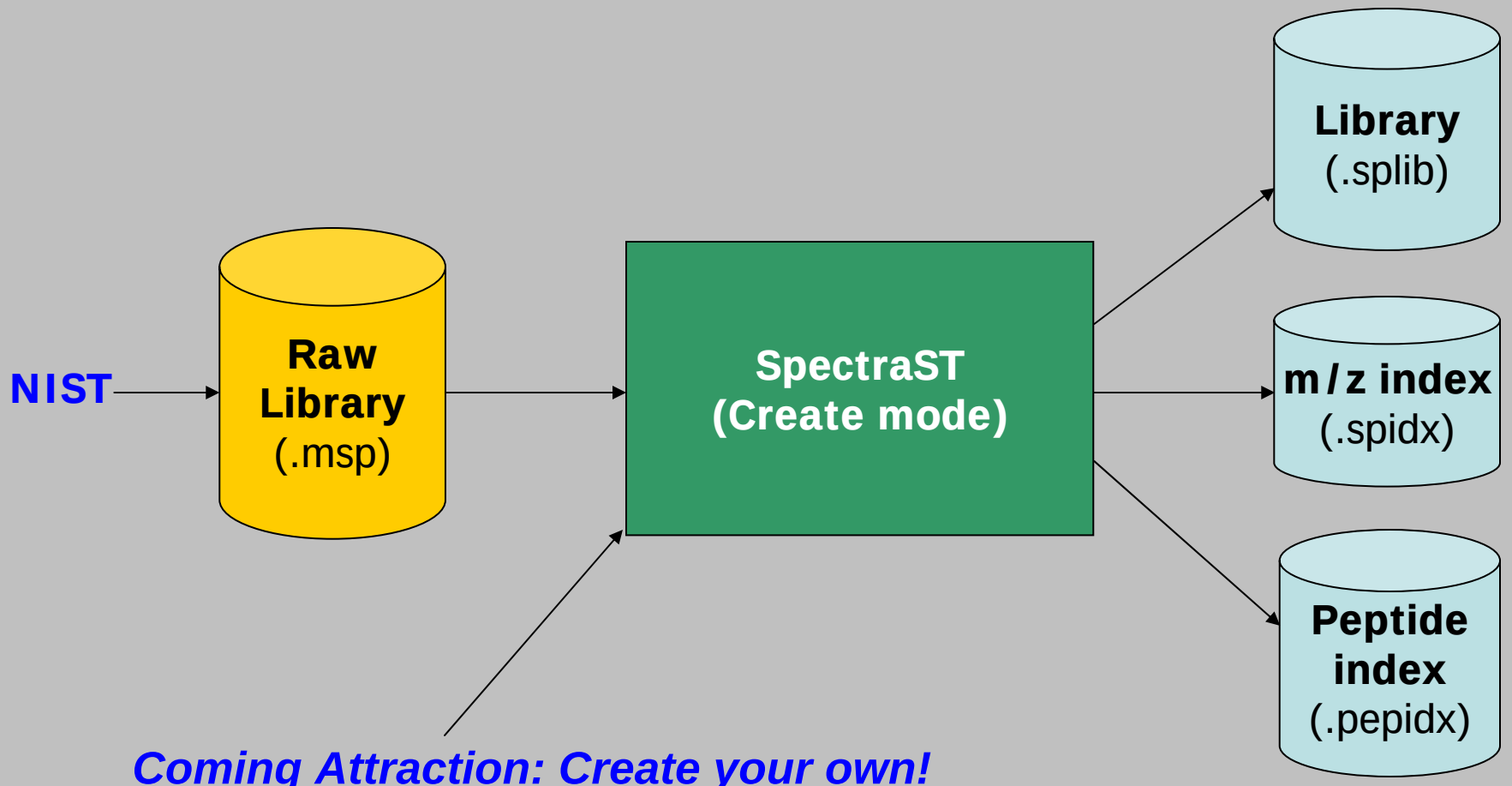
Spectral Searching



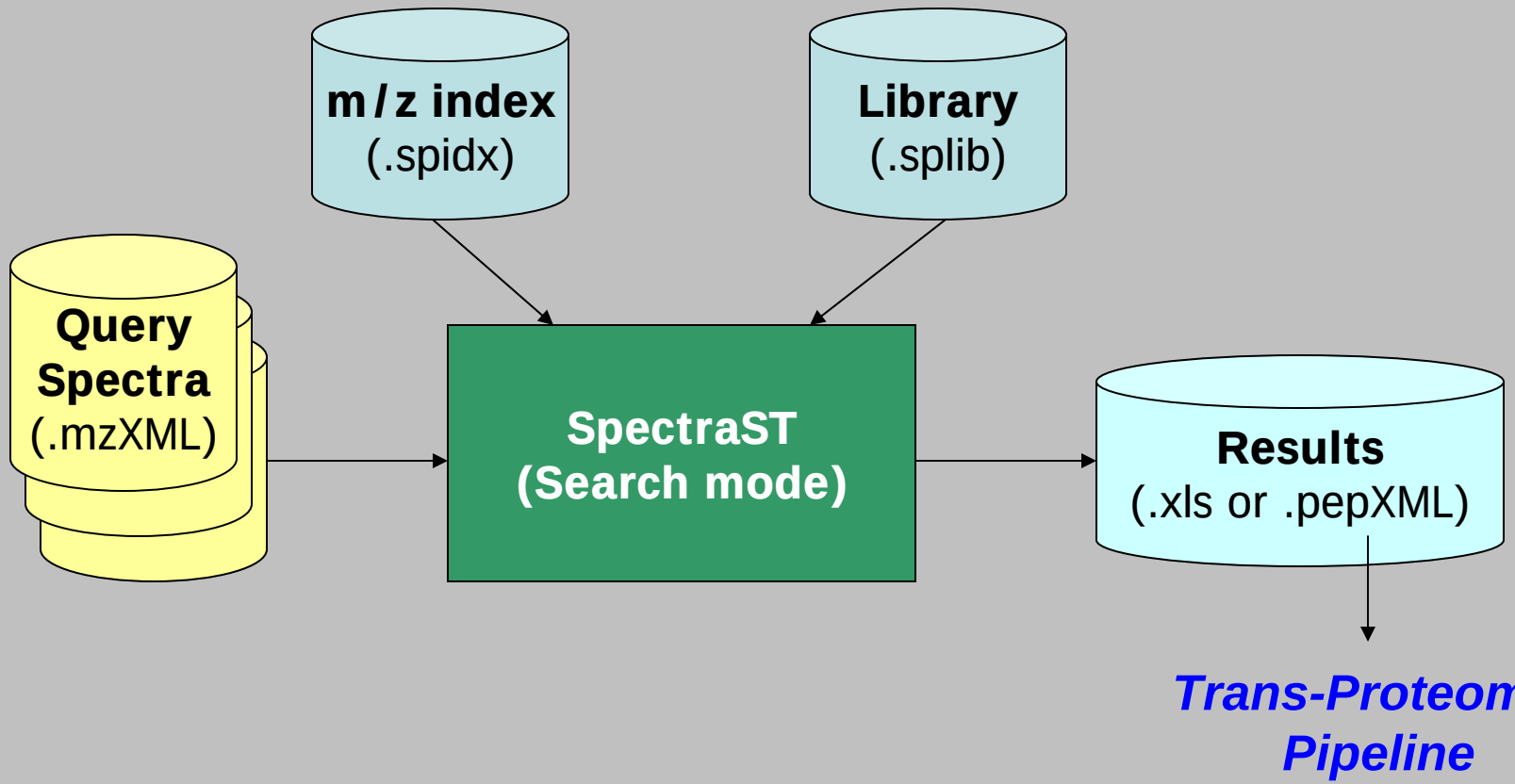
SpectraST

- Open Source (http://sourceforge.net/projects/sashimi under trans_proteomic_pipeline)
- LINUX / Windows versions (http://tools.proteomecenter.org/TPP.php)
- Extensible, modular design
- Fully integrated with Trans-Proteomic Pipeline
- Modest processor and memory requirements

SpectraST *Create* Mode



SpectraST *Search* Mode



SpectraST: Under the Hood

- *Most behavior customizable*
- Query spectra filtering and processing
  - Ignore spectra with too few peaks
  - Remove tiny (noise) peaks
  - Remove the parent peak and its neutral losses
  - Scale intensities (to deemphasize dominant peaks)
    - Scaled intensity = $(\text{intensity})^{0.5}$
  - Assign peaks into unit-m/z bins

SpectraST: Under the Hood

- Similarity scoring
  - Dot product

    (j = bin number, I = normalized scaled intensity)

    $$Dot = \sum_{j=1}^{n} I_{query}(j) I_{library}(j)$$

  - Delta Dot

    $$Delta\ Dot = \frac{Dot(1) - Dot(2)}{Dot(1)}$$

  - Dot Bias

    1.0 = one bin accounts for entire dot product

    ~0.0 = all bins contribute equally

    $$Dot\ Bias = \frac{1}{Dot}\sqrt{\sum_{j=1}^{n} I_{query}^{2}(j) I_{library}^{2}(j)}$$

- Discriminant F function
  - Combination of the scores

SpectraST: Test Drive

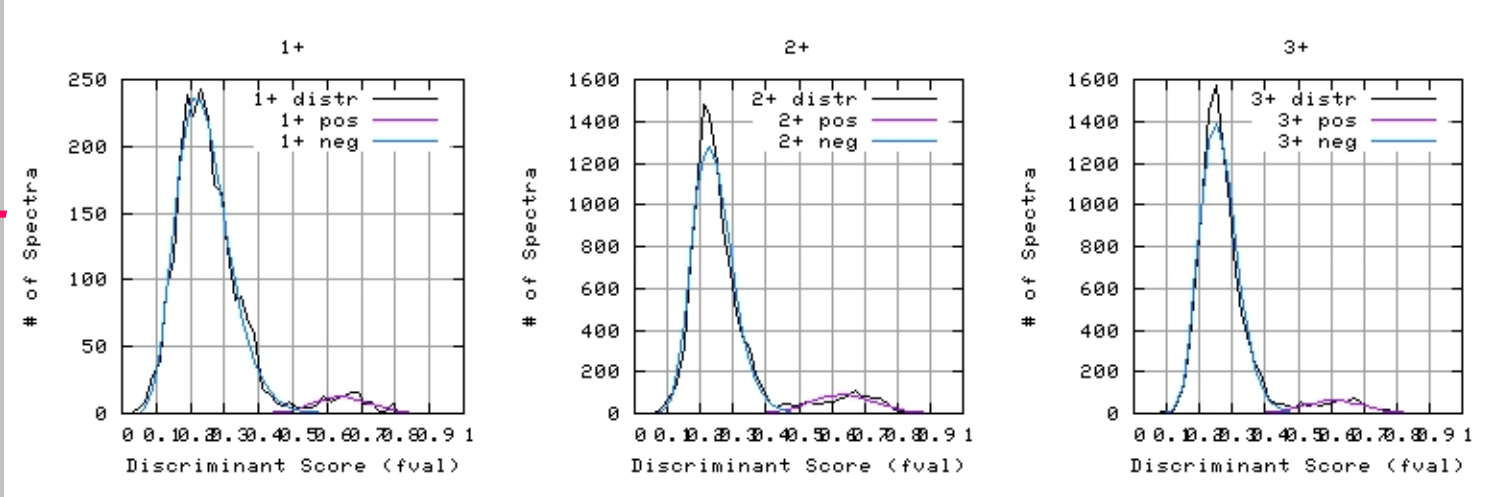
- Test datasets
  - 2 yeast datasets
    - LCQ - ICAT (23,000 spectra; 1,500 p>0.99 IDs)
    - LCQ (2,700 spectra; 990 p>0.99 IDs) *<- Tutorial*
  - 1 human plasma dataset
    - Bruker Esquire (*2.4 million* spectra; 430,000 p>0.99 IDs)

- Speed
  - ~0.01s per query spectrum on P4 3.4GHz, 2GB RAM machine (compared to ~5s for SEQUEST)
  - *~500x improvement in speed!*

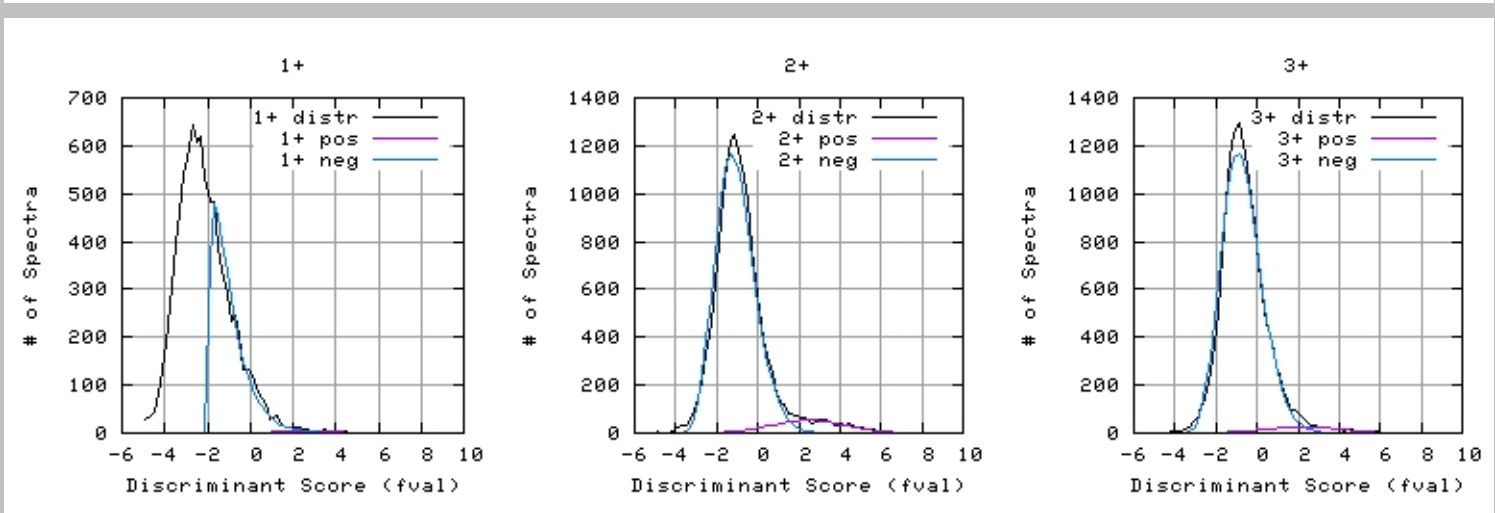
PeptideProphet Analysis

- Yeast dataset 1: LCQ – ICAT (23,000 spectra)

SpectraST



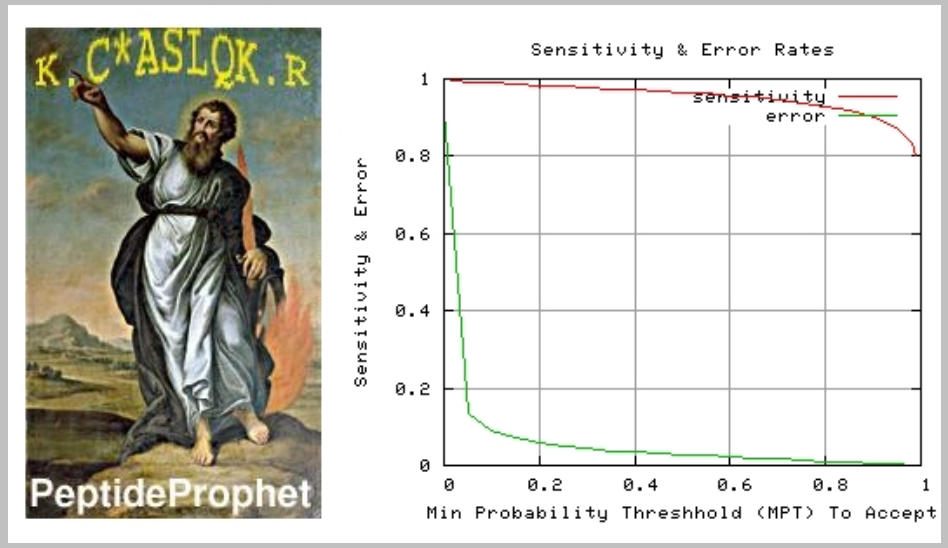
SEQUEST



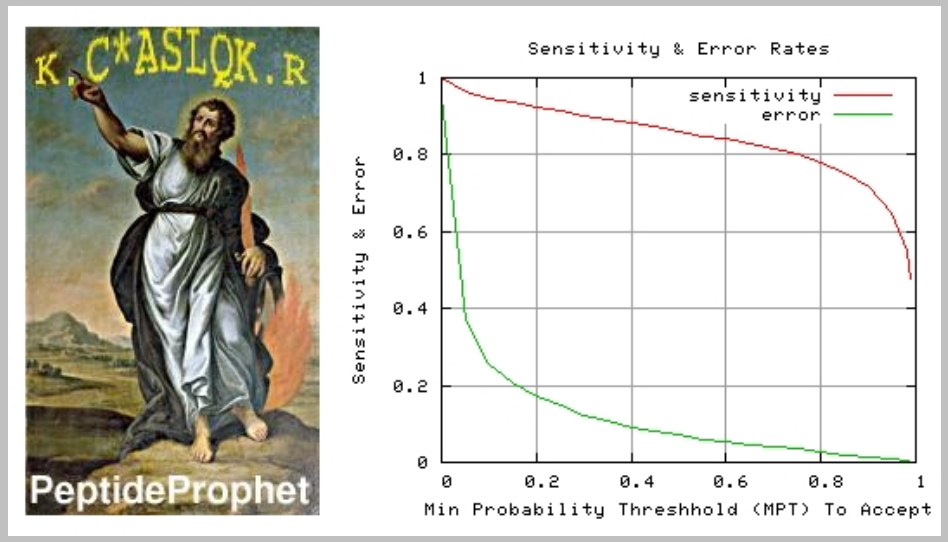
PeptideProphet Analysis

- Yeast dataset 1:
  LCQ – ICAT
  (23,000 spectra)

SpectraST



SEQUEST



Comparison of IDs

		SpectraST	*SEQUEST*
Positive hits (P > 0.99 IDs)		**1,551** *(26% more)*	**1,230**
Intersection (matched positive hits)		**913** *(59%)*	**913** *(74%)*
Extra hits (not in intersection):			
Matched lower-confidence hits		**376** *(24%)*	**78** *(6%)*
Missed by *SpectraST*	**Not In spectral library**		**236** *(19%)*
	In spectral library		**3** *(0.2%)*
Missed by *SEQUEST*	**Manually determined to be correct**	**252** *(16%)*	
	Manually determined to be incorrect	**7** *(0.5%)*	

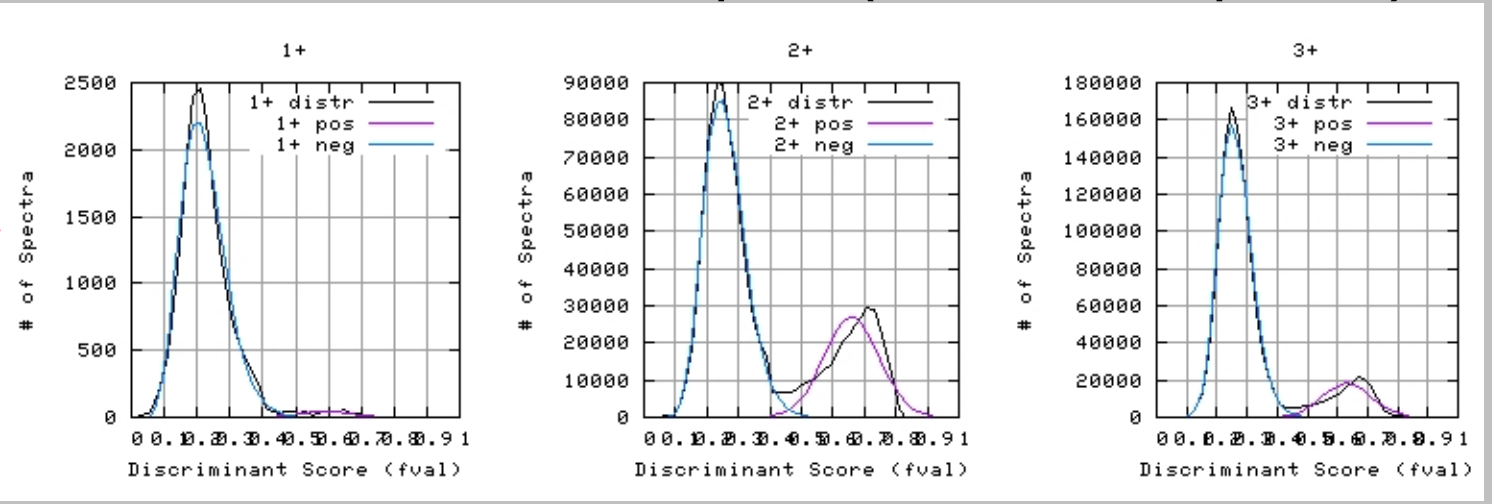
Comparison of IDs

- Lessons
  - SpectraST confidently identifies significantly more spectra than SEQUEST.
  - For IDs found by both engines, SpectraST is more confident about them.
  - When SpectraST misses a good ID, it is almost always because the peptide is not in the library.
  - SpectraST is rarely wrong with its confident IDs.

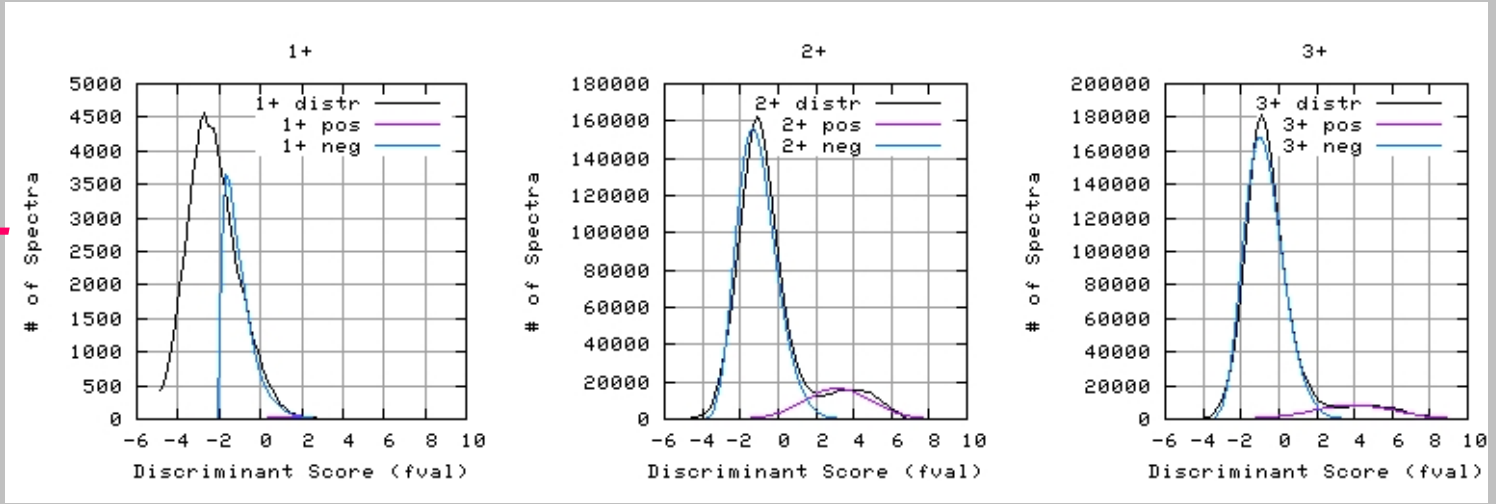
PeptideProphet Analysis

- Human plasma dataset: Bruker Esquire (2.4 million spectra)

SpectraST



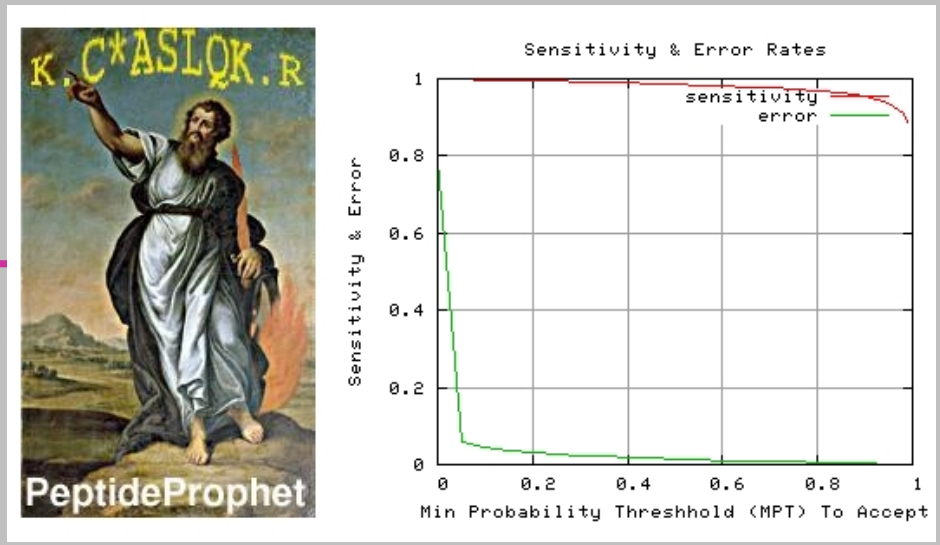
SEQUEST



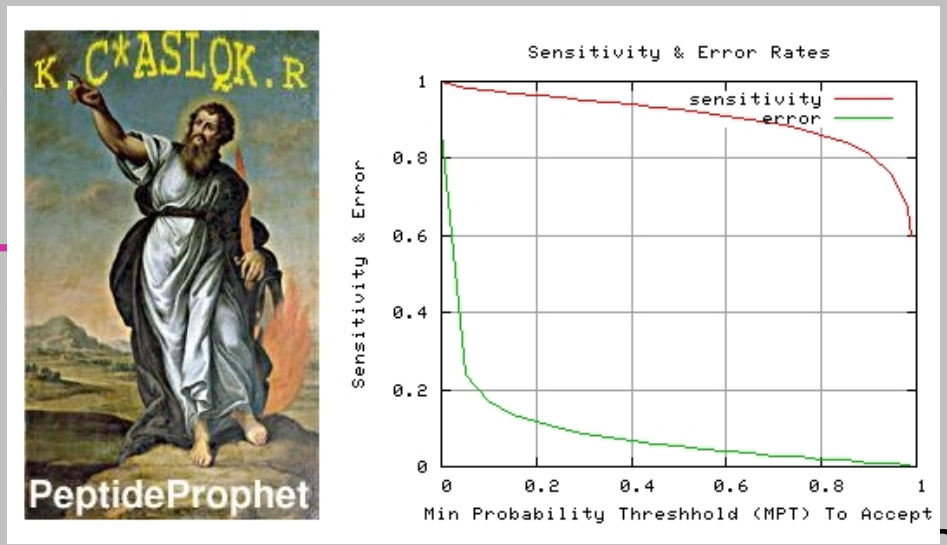
PeptideProphet Analysis

- Human plasma dataset:
  Bruker Esquire
  (2.4 million spectra)

SpectraST



SEQUEST



Comparison of IDs

		SpectraST	*SEQUEST*
Positive hits (P > 0.99 IDs)		**427,056** *(22% more)*	**349,795**
Intersection (matched positive hits)		**177,551** *(41%)*	**177,551** *(51%)*
Extra hits (not in intersection):			
Matched lower-confidence hits		**82,437** *(19%)*	**15,733** *(4%)*
Not matched, but presence of peptide confirmed		**167,557** *(39%)*	**97,918** *(28%)*
Missed by *SpectraST*	**Not In spectral library**		**58,316** *(17%)*
	In spectral library		**277** *(0.08%)*
Missed by *SEQUEST*	**Manually determined to be correct**	**3,836** *(1%)*	
	Manually determined to be incorrect	**39** *(0.01%)*	

Advantages of Spectral Searching (1)

- ***Smaller search space***
  - Only searching against peptides that are known to occur in shotgun proteomics – not putative proteolytic peptides of the entire sequence, most of which will never be observed experimentally
  - No need to search multiple charge states

  which leads to…

  - Huge savings in time
  - More confidence in peptide ID if made multiple times in the past

Search Space

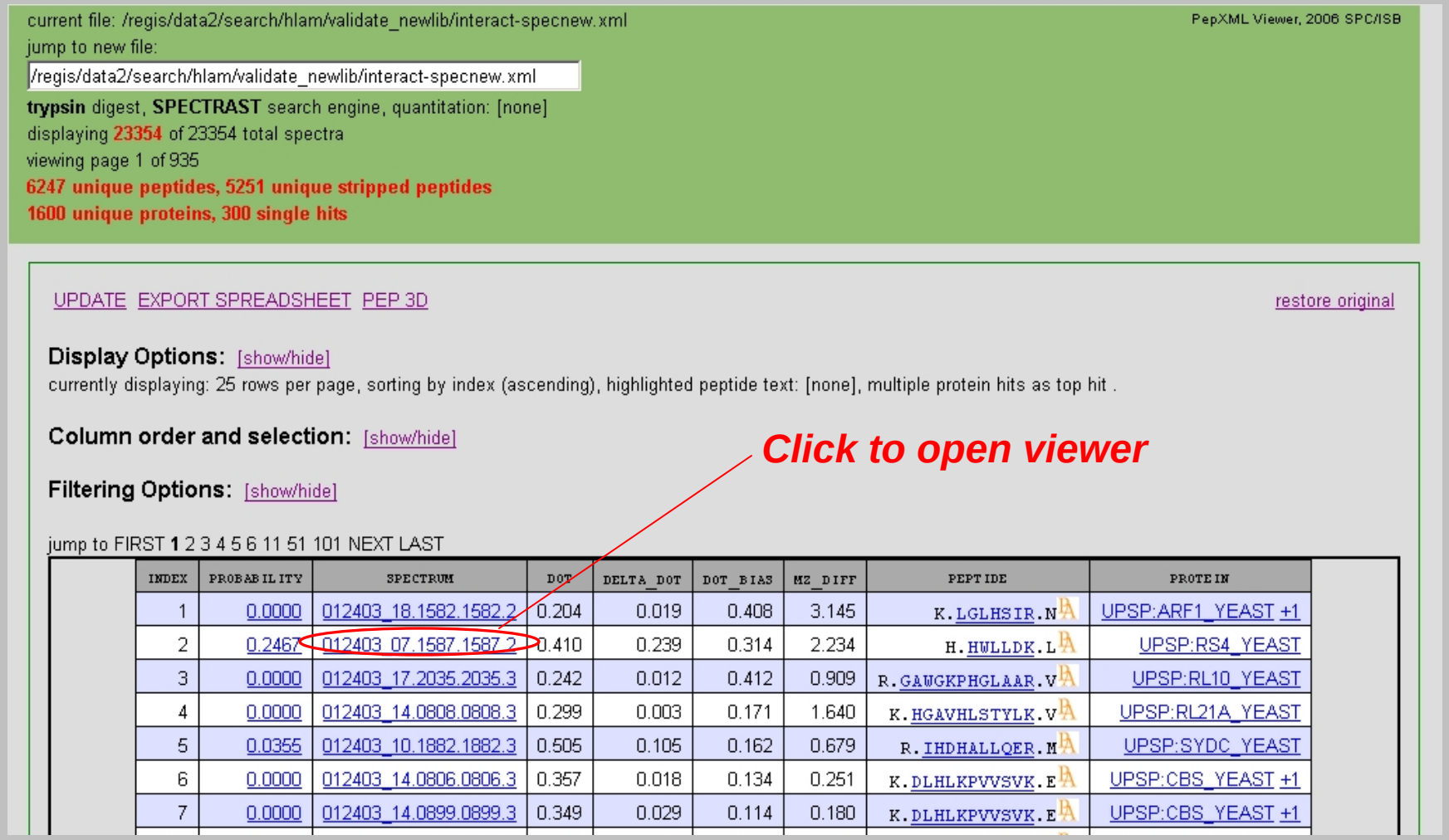
Criterion		Search space	
Sequence characteristics	Precursor m/z range	Sequence Search	Spectral Search
Tryptic, with no missed internal cleavage, +2/+3	300 – 2000	**3.5×10^5**	**1.1×10^4**
	800 – 803	**7.2×10^2**	**52**
At least semi-tryptic, with at most 1 missed internal cleavage, +2/+3	300 – 2000	**1.5×10^7**	**1.7×10^4**
	800 – 803	**3.6×10^4**	**77**

Advantages of Spectral Searching (2)

- ***More precise similarity scoring***
  - Made use of *global similarity* of spectra
  - Peak intensities are accounted for naturally
  - All consistently observed peaks – attributable to common ions or not – are used

  which leads to…
  - Simpler similarity scoring function, greater speed
  - Greater separation between good and bad matches
  - Superior sensitivity and error rates
  - Increased ability to pick up more obscure matches

Spectra Viewer



current file: /regis/data2/search/hlam/validate_newlib/interact-specnew.xml

PepXML Viewer, 2006 SPC/ISB

jump to new file:

/regis/data2/search/hlam/validate_newlib/interact-specnew.xml

trypsin digest, **SPECTRAST** search engine, quantitation: [none]

displaying **23354** of 23354 total spectra

viewing page 1 of 935

6247 unique peptides, 5251 unique stripped peptides

1600 unique proteins, 300 single hits

UPDATE EXPORT SPREADSHEET PEP 3D

restore original

Display Options: [show/hide]

currently displaying: 25 rows per page, sorting by index (ascending), highlighted peptide text: [none], multiple protein hits as top hit .

Column order and selection: [show/hide]

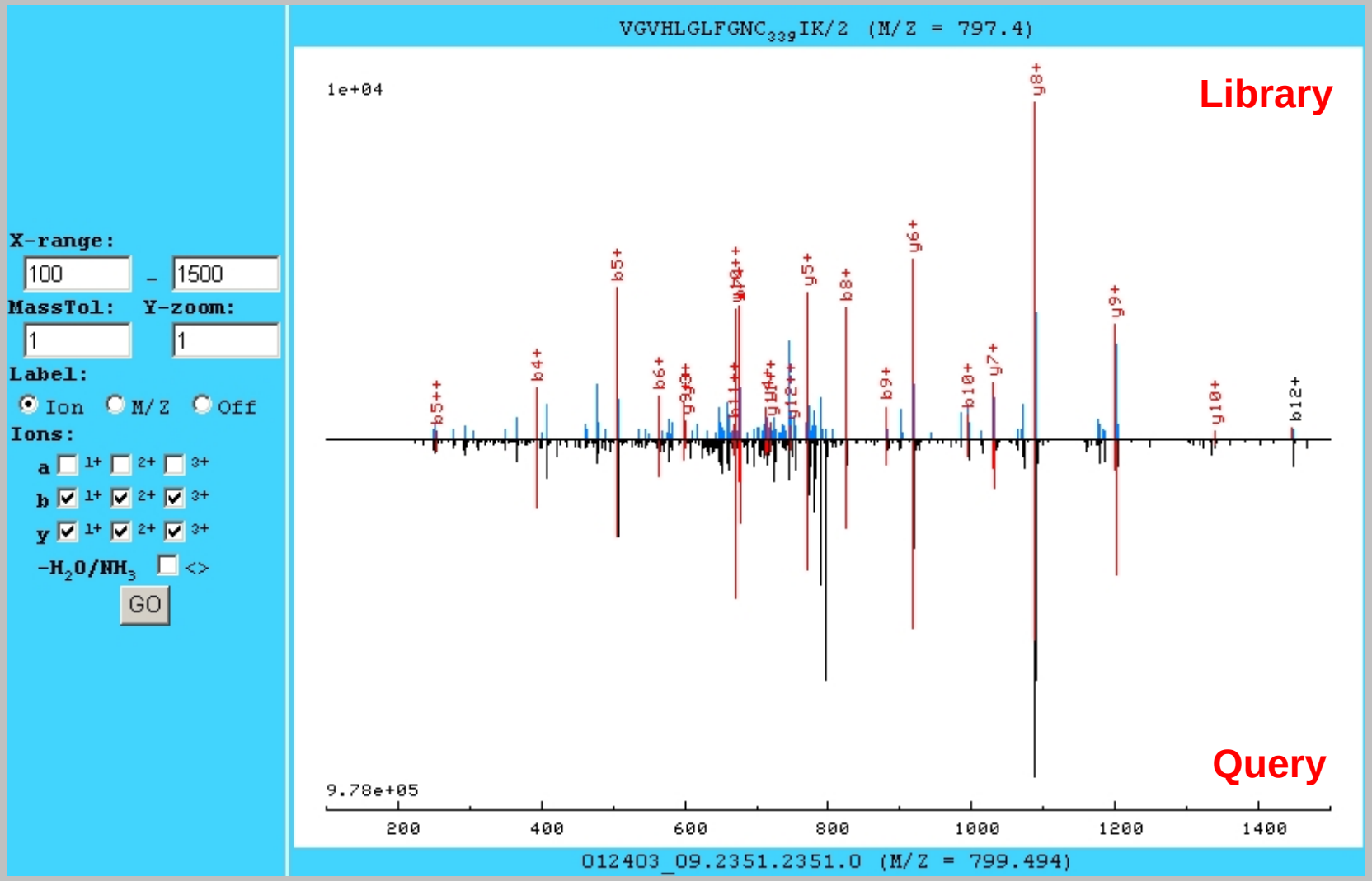
Filtering Options: [show/hide]

Click to open viewer

jump to FIRST **1** 2 3 4 5 6 11 51 101 NEXT LAST

INDEX	PROBABILITY	SPECTRUM	DOT	DELTA_DOT	DOT_BIAS	MZ_DIFF	PEPTIDE	PROTEIN
1	0.0000	012403_18.1582.1582.2	0.204	0.019	0.408	3.145	K.LGLHSIR.N	UPSP:ARF1_YEAST +1
2	0.2467	012403_07.1587.1587.2	0.410	0.239	0.314	2.234	H.HWLLDK.L	UPSP:RS4_YEAST
3	0.0000	012403_17.2035.2035.3	0.242	0.012	0.412	0.909	R.GAWGKPHGLAAR.V	UPSP:RL10_YEAST
4	0.0000	012403_14.0808.0808.3	0.299	0.003	0.171	1.640	K.HGAVHLSTYLK.V	UPSP:RL21A_YEAST
5	0.0355	012403_10.1882.1882.3	0.505	0.105	0.162	0.679	R.IHDHALLQER.M	UPSP:SYDC_YEAST
6	0.0000	012403_14.0806.0806.3	0.357	0.018	0.134	0.251	K.DLHLKPVVSVK.E	UPSP:CBS_YEAST +1
7	0.0000	012403_14.0899.0899.3	0.349	0.029	0.114	0.180	K.DLHLKPVVSVK.E	UPSP:CBS_YEAST +1

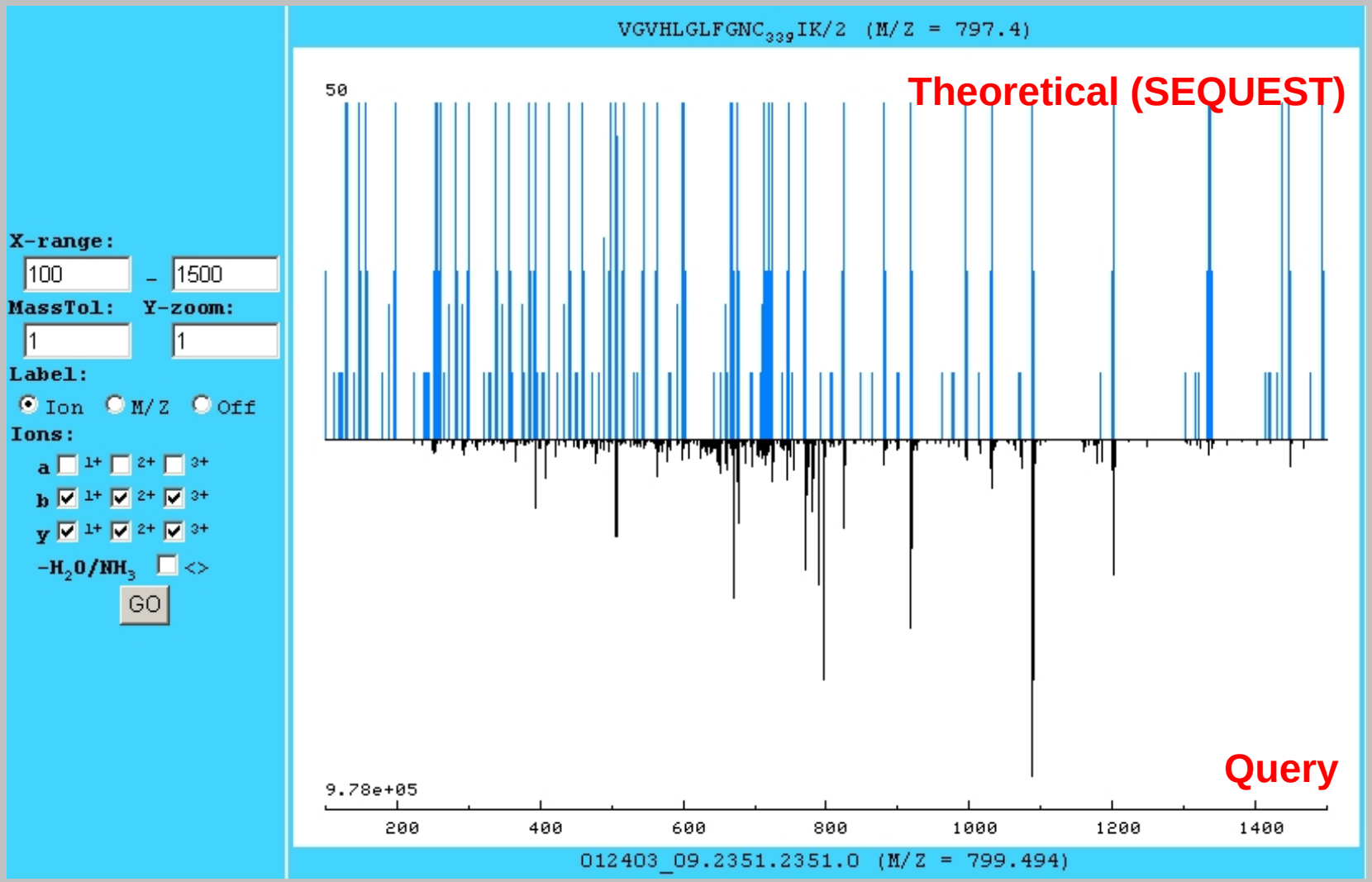
Spectral Match



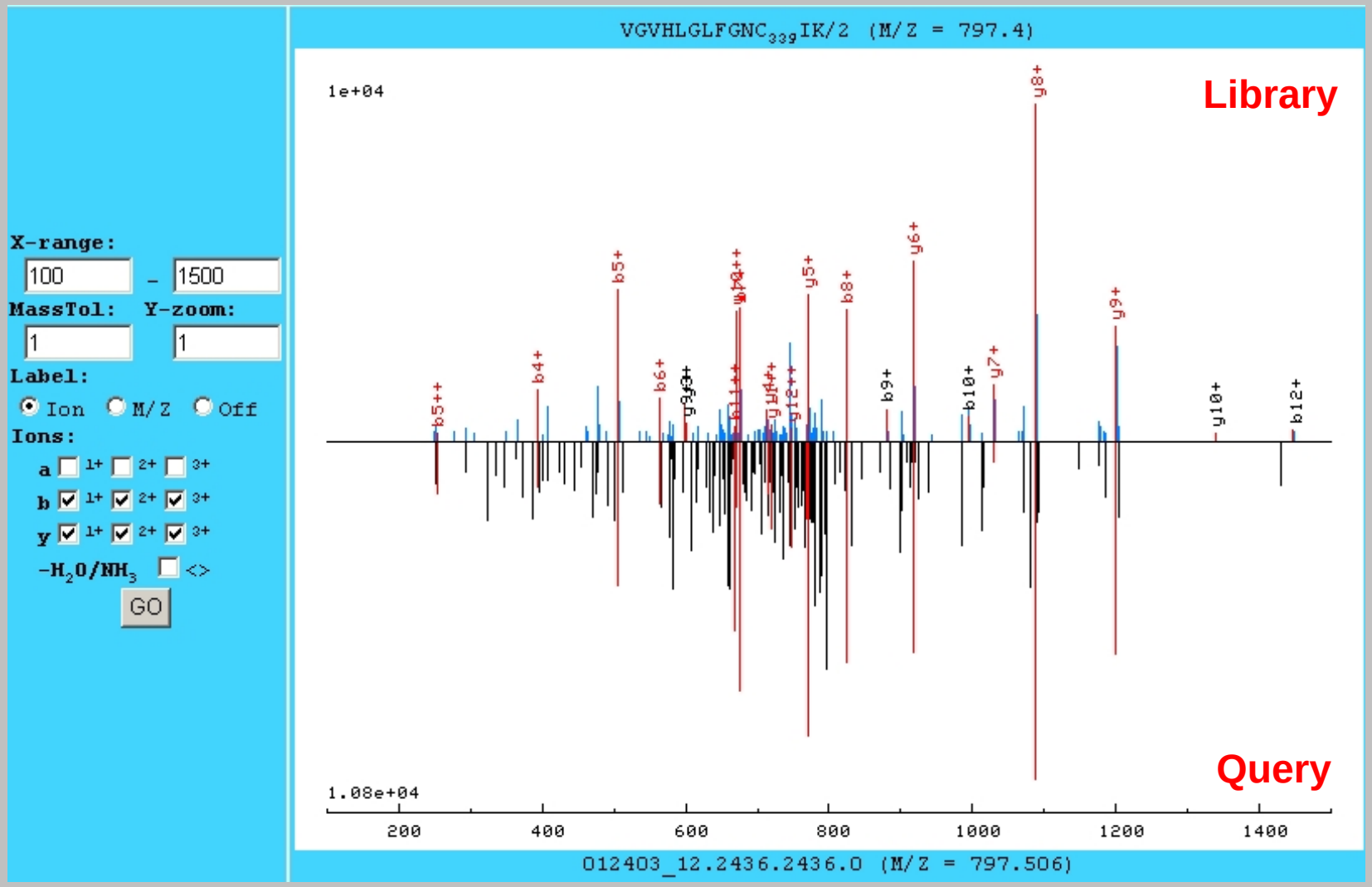
Spectral Match

b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
100.1400	50.5737	1	V	13		
157.1919	79.0996	2	G	12	1494.7177	**747.8626**
256.3245	128.6659	3	V	11	1437.6658	**719.3366**
393.4656	197.2365	4	H	10	**1338.5332**	**669.7703**
506.6250	**253.8162**	5	L	9	**1201.3921**	**601.1998**
563.6769	282.3421	6	G	8	**1088.2327**	544.6201
676.8363	338.9218	7	L	7	**1031.1808**	516.0941
824.0129	412.5101	8	F	6	**918.0214**	459.5144
881.0648	441.0361	9	G	5	**770.8448**	385.9261
995.1686	498.0880	10	N	4	**713.7929**	357.4002
1334.5016	**667.7545**	11	C[339]	3	**599.6891**	300.3483
1447.6610	724.3342	12	I	2	260.3561	130.6818
		13	K	1	147.1967	74.1021

Theoretical Spectrum



Noisy Query Spectrum



Noisy Query Spectrum

b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
100.1400	50.5737	1	V	13		
157.1919	79.0996	2	G	12	1494.7177	**747.8626**
256.3245	128.6659	3	V	11	1437.6658	**719.3366**
393.4656	197.2365	4	H	10	**1338.5332**	**669.7703**
506.6250	**253.8162**	5	L	9	**1201.3921**	**601.1998**
563.6769	282.3421	6	G	8	**1088.2327**	544.6201
676.8363	338.9218	7	L	7	**1031.1808**	516.0941
824.0129	412.5101	8	F	6	**918.0214**	459.5144
881.0648	441.0361	9	G	5	**770.8448**	385.9261
995.1686	498.0880	10	N	4	**713.7929**	357.4002
1334.5016	**667.7545**	11	C[339]	3	**599.6891**	300.3483
1447.6610	724.3342	12	I	2	260.3561	130.6818
		13	K	1	147.1967	74.1021

Advantages of Spectral Searching (3)

- ***Implicitly searching multiple sequence search engines***
  - Sequence search engines have their own idiosyncrasies, and will yield largely overlapping (~70% typical), but not identical set of peptide identifications
  - NIST spectral libraries are compiled by combining the confident hits of 4 different search engines
  - Spectral searching take advantage of the strength of each of the engines without the additional time and effort

  which leads to…

  - More identifications
  - Additional confidence for identifications found by multiple search engines

Potential Pitfalls

- Library coverage
  - Will only find peptides that are previously observed and represented in the library
    - OK for targeted proteomics: you know what you are looking for
  - Should improve over time with more data
- Library quality
  - A poorly constructed library will lead to false positives and negatives
    - Mis-identified library spectra
    - Noisy or impure library spectra
    - Similar spectra mapping to distinct peptides
  - Need stringent confidence thresholds and quality filters

Future Outlook

- Libraries
  - More organisms (Drosophila, mouse)
  - Coverage will improve with more data *(Please contribute!)*
    - More sampling conditions
    - More advanced instruments
    - More kinds of modifications
  - Custom library building functionalities in SpectraST future release – *build your own libraries* for your favorite organism/tissue/proteins
- Searching
  - Web interface on PeptideAtlas
  - Improvement in SpectraST workflow and scoring

Conclusions

- *SpectraST* is an open-source tool for spectral library searching of peptide MS/MS spectra
  - Fast
  - Improved sensitivity and error rate
  - More identifications
  - Integrated into TPP
- *Spectral library searching is here and will get better*
  - Applicable for non-discovery type experiments
  - Will eventually replace sequence searching in most typical shotgun proteomics workflows

Tutorial