

Annotation of the Yeast Proteome with PeptideAtlas

Nichole L. King¹, Eric W. Deutsch¹, Jeffrey A. Ranish¹, Alexey I. Nesvizhskii², James S. Eddes¹, Parag Mallick^{3,4}, Frank Desiere⁵, Mark Flory⁶, Daniel B. Martin⁷, Jimmy Eng^{1,8}, Bong Kim¹, Hookeun Lee⁹, Brian Raught¹⁰, Ruedi Aebersold^{1,9}

¹Institute for Systems Biology, 1441 N 34th Street, Seattle, WA 98103, USA; ²Dept of Pathology, University of Michigan, M5226 Med Sci I, 1301 Catherine Road, USA; ³Cedars-Sinai Medical Center; ⁴Louis Warschaw Prostate Cancer Center; ⁵Nestlé Research Center, 1000 Lausanne 26, Switzerland; ⁶Dept of Molecular Biology and Biochemistry, Wesleyan University, Middletown, CT 06459; ⁷Divisions of Human Biology and Clinical Research, Fred Hutchinson Cancer Research Center Seattle WA 98109-1024; ⁸PHSD, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; ⁹IMSB, Swiss Federal Institute of Technology, ETH Honggerberg, Zurich, Switzerland; ¹⁰University Health Network, Ontario Cancer Institute and McLaughlin Centre for Molecular Medicine, MaRS TMDT 9-805, 101 College Street, Toronto, ON M5G 1L7, Canada

Overview:

We demonstrate that MS/MS spectra from a large number of diverse sources can be uniformly processed and combined to create a large dataset useful for exploring the *S. cerevisiae* proteome. The public interfaces can be useful for designing new experiments, and the datasets can be used for proteomic software development.

Introduction:

We present the *S. cerevisiae* PeptideAtlas, a compendium of peptides from 45 diverse experiments and 4.1 million MS/MS. The observed peptides align to ~75% of named yeast protein coding genes, 50% of the uncharacterized ORFs in the Saccharomyces Genome Database (SGD), and 52% of yeast ORFs with a Gene Ontology annotation of "molecular function unknown". The *S. cerevisiae* PeptideAtlas thus possesses the highest degree of proteome coverage for any eukaryotic organism to date in a single public database offering entire datasets as validation. Here, we highlight the resources that we will make available, and demonstrate how the *S. cerevisiae* PeptideAtlas may be used for data mining. For example, the Atlas may be used to construct high quality lists of observable peptides for synthesis and targeted proteomics, to validate predicted proteins, and to serve as a resource for software development. This large compendium of observed peptides thus represents a novel and important tool for the study of yeast proteins. Importantly, the *S. cerevisiae* PeptideAtlas will grow and improve as researchers continue to contribute additional experimental data.

Methods:

The yeast PeptideAtlas is comprised of tandem mass spectrometry datasets originating from a variety of different labs, sample treatment protocols, separation protocols, and mass spectrometers. The spectra were converted to mzXML format, searched with SEQUEST against a large database of yeast proteins, and assigned probabilities of being correct using the PeptideProphet implemented in the Trans-Proteomic pipeline [see Figure 1]. The results were stored in the SBEAMS Proteomics database. The resulting peptide information above an identification threshold of P>0.9 was gathered into a group, aligned against the SGD database using BLAST, and assigned calculated chromosomal coordinates. The contents were stored with public and private access restrictions in the SBEAMS PeptideAtlas database and raw data repository, and integrated into the framework of our web interfaces.

Results:

• The cumulative number of MS/MS versus unique peptide identifications continues to rise [see Figure 2]

• There is a mass bias in the peptides we are observing [see Figure 3] and a hydrophobicity bias (not shown).

• Summary statistics of the current yeast PeptideAtlas build are presented in Table 1. The number of distinct peptide sequences identified in these spectra (with P>0.9) was 35,019. The number of SGD protein coding genes observed was 75%.

• The numbers of observed ORFs categorized by SGD annotation are shown in Table 2. We see 50% of uncharacterized ORFs (homologs not yet observed in Sc).

• In Figure 4 we examined the distribution of observed ORFs and unobserved ORFs using the Ghaemmaghami et al. 2003 codon enrichment correlations and found that the unobserved ORFs are likely composed bonafide ORFs as well as ORFs that are not likely to code for proteins.

• The proteins identified in the yeast PeptideAtlas are generally evenly distributed with respect to populated Gene Ontology (GO) terms [see Figure 5]. Interestingly, we observe a large percentage of ORFs annotated as "molecular function unknown".

These findings are part of a manuscript in preparation and so the yeast organism database is not yet publicly available.

Conclusions:

• We have demonstrated that MS/MS spectra from a large number of diverse sources can be uniformly processed and combined to create a large dataset useful for exploring the yeast proteome.

• The web interfaces can be useful for designing new experiments (visit www.peptideatlas.org to search the databases)

• The YeastPeptideAtlas can be used for other proteomic software development. For examples, please see the following:

- A nearby poster "SpectraST: An Open-Source MS/MS Spectra-Matching Library Search Tool for Targeted Proteomics" (slot 530) by Henry Lam et al.
- A paper in progress "A data set of high quality unassigned tandem mass spectrometry spectra" by James Eddes et al.
- A paper in progress "A Computational Approach for Identifying and Predicting Proteotypic Peptides for Quantitative Proteomic Experiments" by Parag Mallick et al.

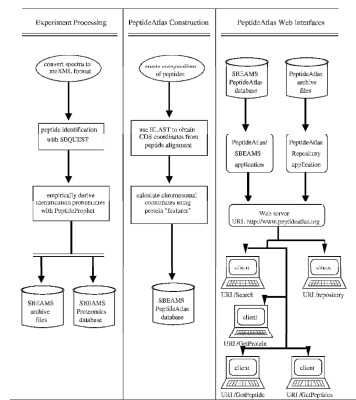


Figure 1: PeptideAtlas processing, creation, and interfaces. The first column outlines experiment level processing, the second column shows major stages in the construction of a PeptideAtlas, and the third column shows the server/client model providing web interfaces to PeptideAtlas.

www.peptideatlas.org

Acknowledgements

We thank NHLBI for partial funding of this project under contract No. N01-HV-28179. We also thank Olga Vitek and Julian Watts for their advice and consultation. We thank Steve Stein and are grateful to all of the researchers who have made their datasets publicly available, specifically, Marcello Marelli and collaborators, Peng Lu at OPD, P. Haynes and collaborators, K.R. Serikawa and collaborators, S. Gygi and collaborators, Ho and collaborators, and Trey Ideker.

Number of Experiments	45
Number of MS runs	1325
Number of MS/MS searched	4,110,193
Number of MS/MS w/ P>0.9	535,925
Number of distinct peptides w/ P>0.9	35,019
Number of distinct peptides with perfect alignment to SGD	34,345
Number of SGD protein coding genes	4465
Number of SGD protein coding genes seen in PeptideAtlas	3329 (= 75 %)

Table 1: PeptideAtlas characteristics. The percentage of SGD protein coding genes seen in PeptideAtlas is 75% if we include all P>0.9 peptides, but if we remove from the peptide list those that have been observed only once, 55% of the SGD protein coding genes have been observed in the Atlas.

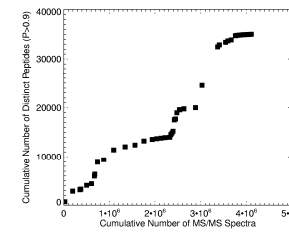


Figure 2: Cumulative number of MS/MS spectra versus the number of unique peptide identifications with P > 0.9. The slope is nearly vertical in regions of the curve where similar experiments were performed. The curve is expected to show saturation when additional spectra provide no new peptides above P>0.9.

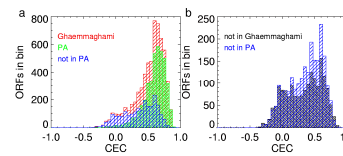


Figure 4: Histogram of codon enrichment correlations (CECs) for (a) all ORFs listed in Ghaemmaghami et al. 2003, for all ORFs seen in PeptideAtlas, and for all ORFs not seen in PeptideAtlas; and (b) histogram of CEC for ORFs not seen in the Ghaemmaghami et al. expression sets and for all ORFs not seen in PeptideAtlas. Observed proteins in PeptideAtlas and in Ghaemmaghami et al. show a positive skew in CEC showing that they deviate significantly from that expected from random sequence of codons, while the unobserved proteins show more uniformly distributed CEC values.

ORF annotation	PeptideAtlas	SGD	%(PA/SGD)
Uncharacterized	718	1423	50%
Verified	3279	4357	75 %
Verified silenced gene	1	4	15 %
Dubious	94	820	11 %
pseudogene	7	21	33 %
Transposable element gene	87	89	98 %
TOTAL	4186	6714	62 %

Table 2: Proteins in SGD ORF annotation categories. The percent of SGD ORFs seen in the PeptideAtlas are shown in column 4. If we remove from the PeptideAtlas those peptides which have only been observed once, the observed percentages of the annotation categories become 29%, 56%, 0%, 1%, 5%, and 15% for uncharacterized, verified, verified/silenced gene, dubious, pseudogene, and transposable element gene, respectively.

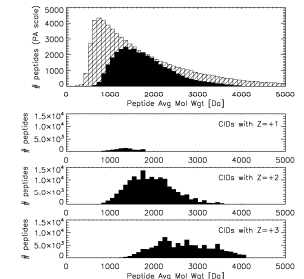


Figure 3: Peptide mass distributions in PeptideAtlas (solid filled bars) and in an in-silico tryptically digested SGD protein database (hashed bars) allowing one missed cleavage. The unique peptide sequences are counted rather than the total number of occurrences. The large number of peptides with masses less than 1000 Da are outside of many mass spec acquisition settings, and are peptides that are difficult to identify in database searches.

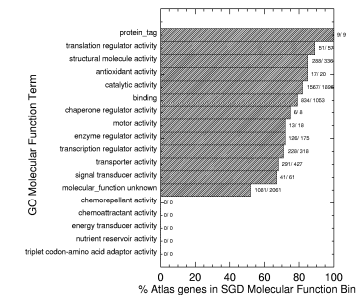


Figure 5: The number of genes identified in GO Molecular Function terms (more specifically, the first level children of the terms). The bars are annotated with the number of SGD genes annotated in that term and the number of SGD Genes seen in PeptideAtlas for that term. Many of the GO terms annotated as "molecular function unknown" are present in the PeptideAtlas.