CPAS Overview

Josh Eckels

LabKey Software

jeckels@labkey.com

FRED HUTCHINSON CANCER RESEARCH CENTER

A LIFE OF SCIENCE

CPAS

- Web-based system for processing, storing, and analyzing results of MS/MS experiments
- Key goals:
  - Provide a great analysis front-end for the TPP tools
  - Handle high-throughput processing and analysis of results
  - Provide universal access to data and support collaboration
  - Keep data private & secure
  - Make it easy to install, administer, and use
  - Allow queries based on experimental protocols and samples
  - Support popular operating systems & database servers
  - Use public file formats for import, export, and exchange
  - Distribute via liberal open source license (Apache 2.0)

Brief CPAS History

- 2003 – 2004
  - Dr. Martin McIntosh's laboratory receives grant from NCI; includes ISB as partner
  - Initial system developed for proteomics research
- 2005
  - CPAS 1.0 product, source code, and publication released
  - Core annotation system (based on FuGE) suitable for generic biological portal
  - LabKey Software formed by FHCRC and former employees to support CPAS
    - Independent consulting company
    - Provides support and service to other institutions

Brief CPAS History

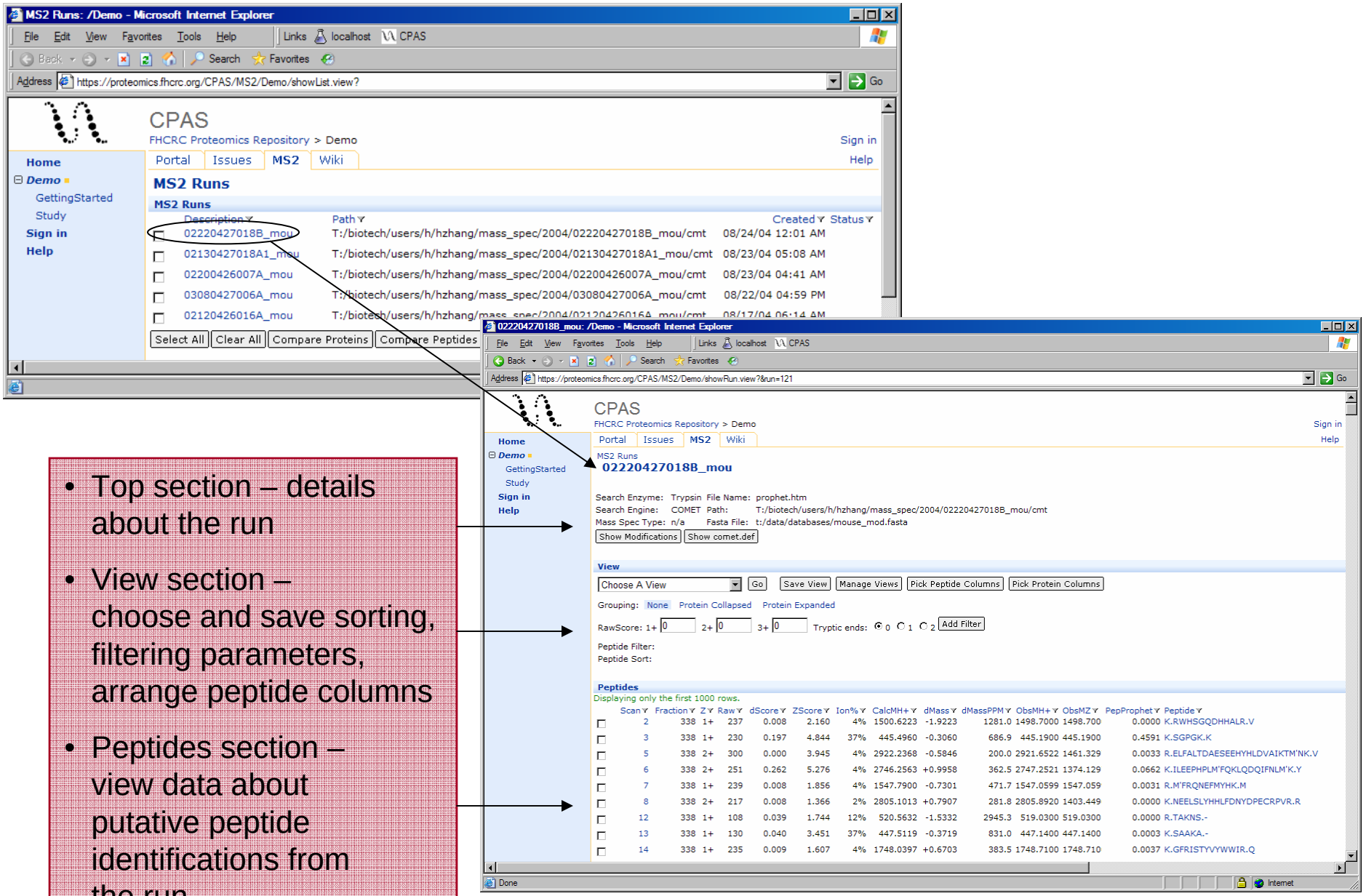
- Traction
  - FHCRC CPAS: 19,000 MS/MS runs containing 180 million peptide ids and spectra
  - Over 200 institutions have downloaded the system
- Developers contributing
  - FHCRC: Driving extensions to proteomics features
  - LabKey: Platform & proteomics dev, other modules (flow cytometry, observational studies)
  - Bioinformatics Institute of Singapore, University of Washington, University of Kentucky, Cedars-Sinai

Key MS/MS Analysis Features

- Load MS/MS results produced by many common search engines
  - Mascot, X! Tandem, SEQUEST, COMET
- Inspect individual MS/MS spectra
- Filter and sort results based on peptide and protein characteristics:
  - Search engine scores, PeptideProphet™, delta mass, modifications, etc.
  - Sequence mass, sequence coverage, gene name, ProteinProphet™ score, etc.
- Group results by protein or ProteinProphet groups
- Customize columns, save favorite filters and views
- Export filtered, sorted results to Excel, TSV, DTA, PKL formats
- Filter groups of runs and compare peptides/proteins between them
- Analyze quantitation of peptides & proteins (XPRESS, Q3, ProteinProphet)
- Link results to rich protein annotations & experimental annotations
- Expose results for programmatic access through caBIG™ interface

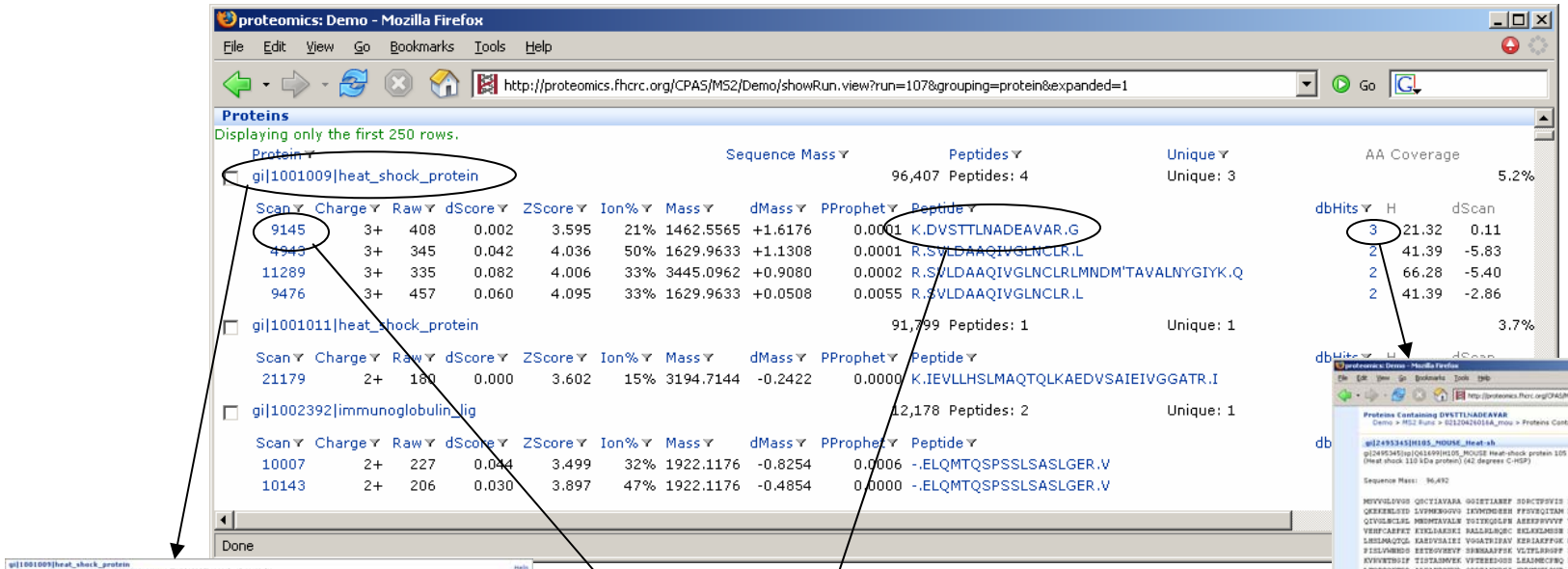
Demo

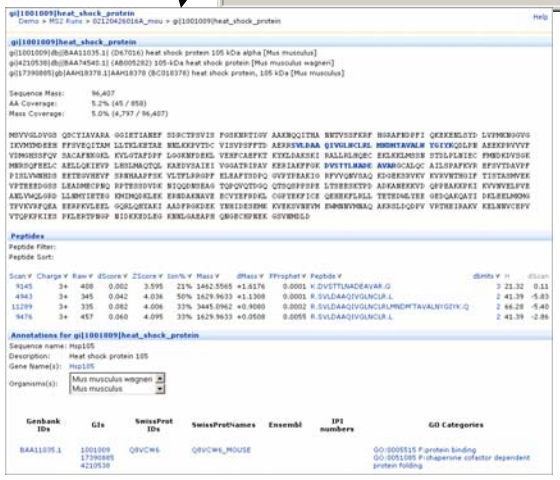
Viewing Runs



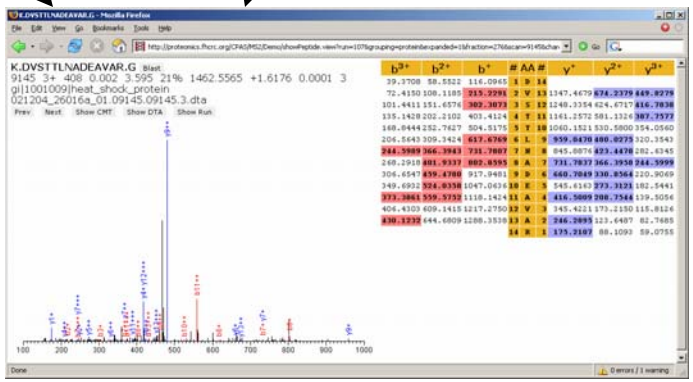
- Top section – details about the run
- View section – choose and save sorting, filtering parameters, arrange peptide columns
- Peptides section – view data about putative peptide identifications from the run

Expanded Protein View

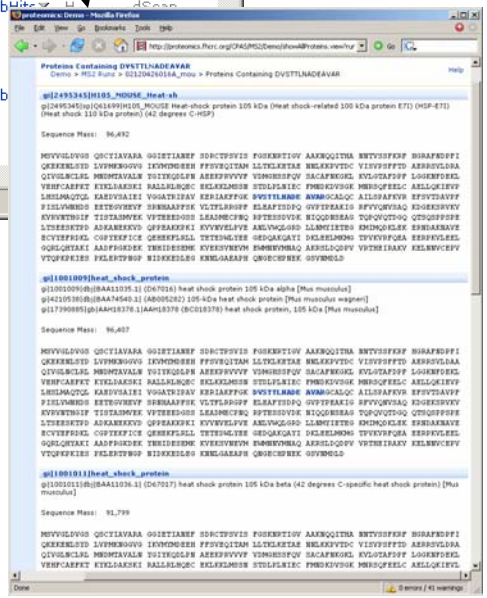




Protein Details



Individual MS/MS spectrum



Protein Hits

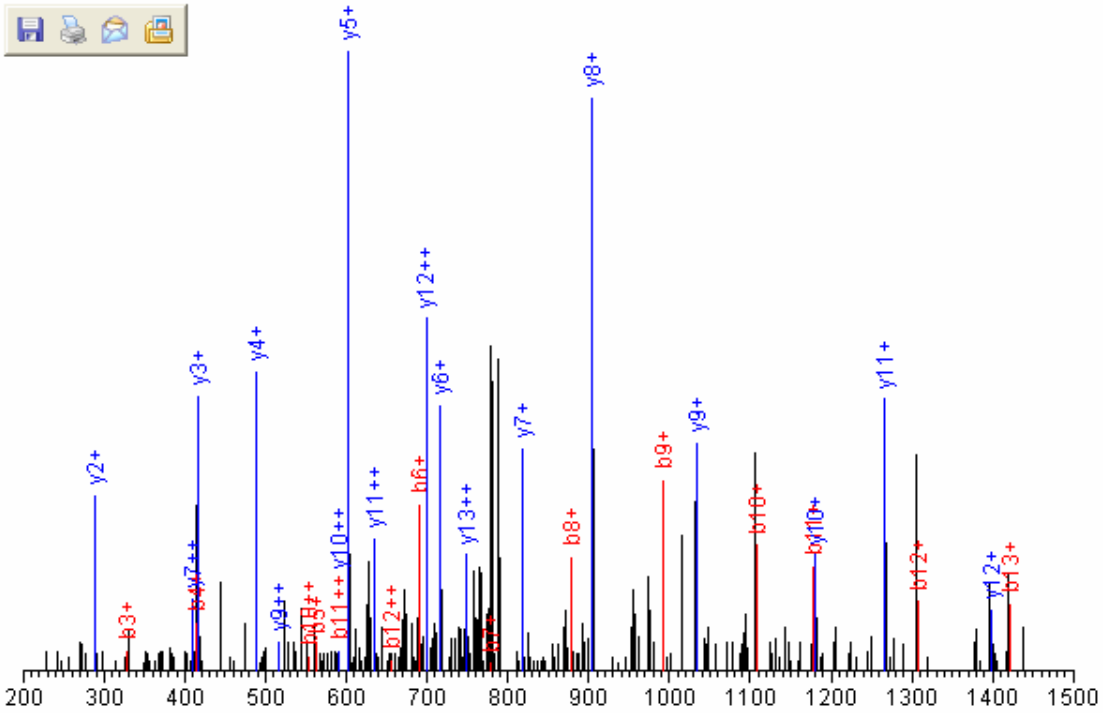
R.VVESM'QSTLDAEIR.S Blast

7045 2+ 642 0.328 0.000 67% 1594.7778 -0.0657 0.9909 7

gi|90095|S00084_myosin_heavy_

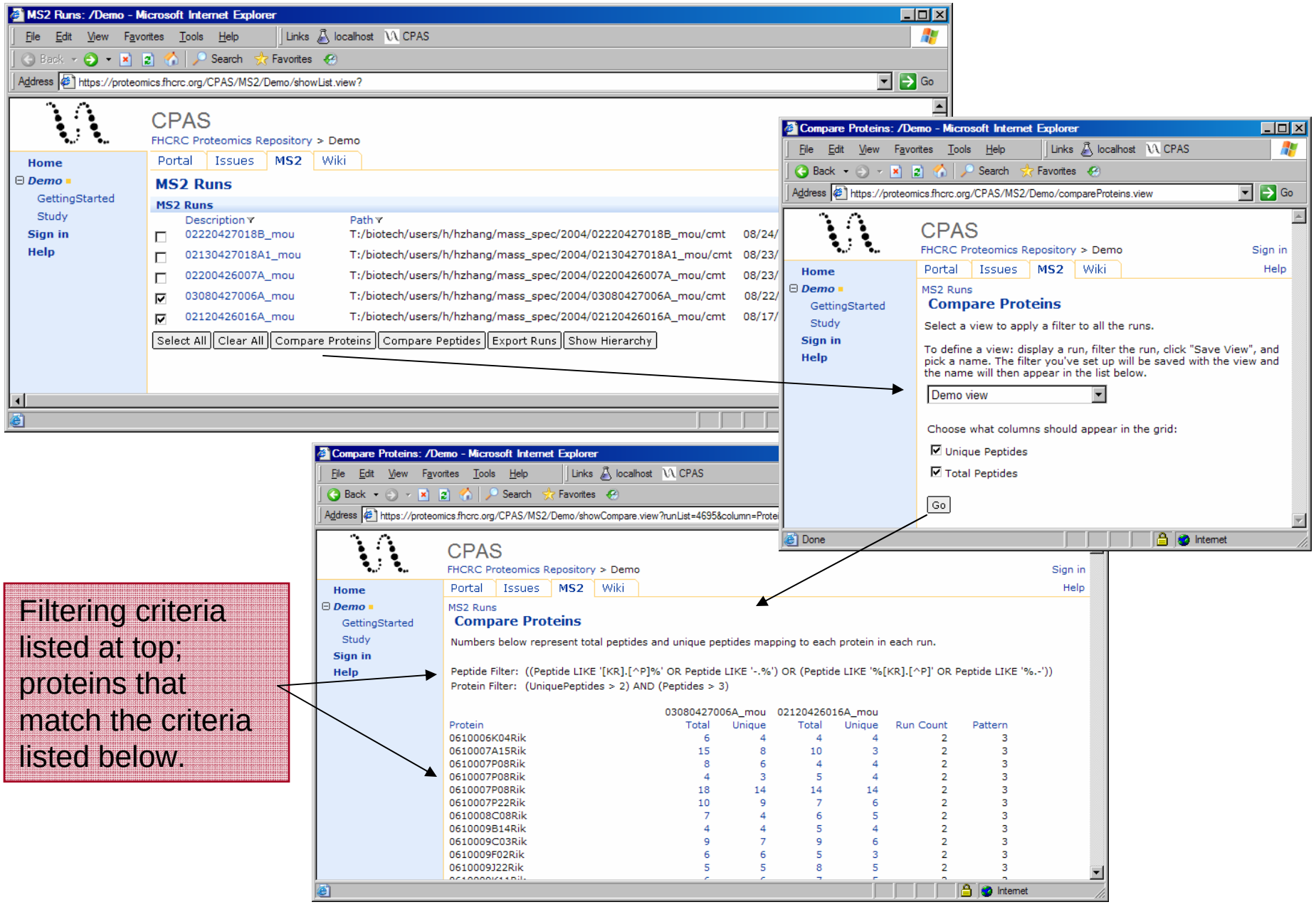
030804_27006A_01.7045.7045.2.dta

Prev Next Show CMT Show DTA Show Run



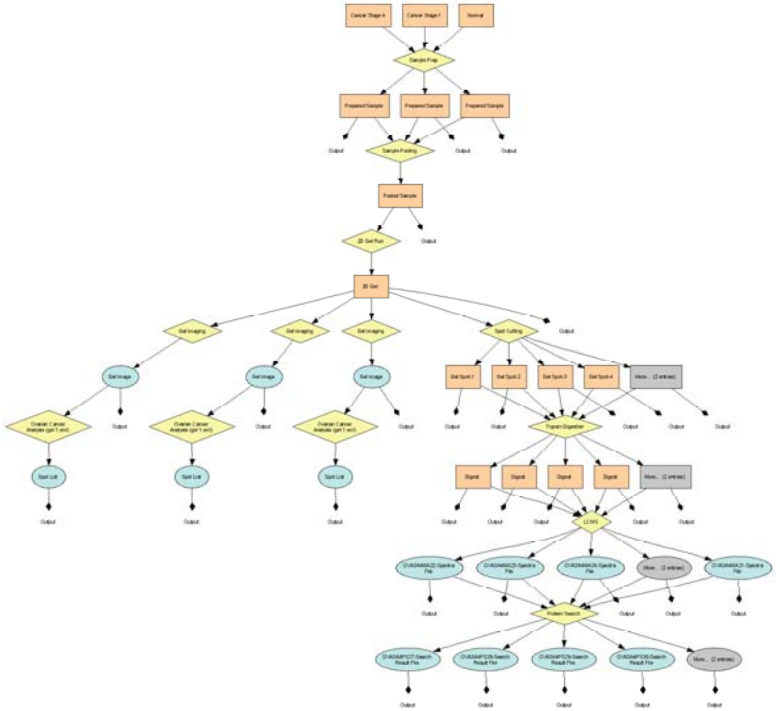
b^{2+}	b^{+}	#	AA	#	y^{+}	y^{2+}
50.5742	100.1405	1	V	14		
100.1405	199.2731	2	V	13	1495.6453	**748.3266**
164.6982	**328.3885**	3	E	12	**1396.5127**	**698.7603**
208.2373	**415.4667**	4	S	11	**1267.3972**	**634.2026**
281.8336	**562.6593**	5	M'	10	**1180.3190**	**590.6635**
345.8990	**690.7900**	6	Q	9	**1033.1265**	**517.0672**
389.4381	**777.8682**	7	S	8	**904.9957**	453.0018
439.9906	**878.9733**	8	T	7	**817.9175**	**409.4627**
496.5703	**992.1327**	9	L	6	**716.8125**	358.9102
554.1146	**1107.2213**	10	D	5	**603.6530**	302.3305
589.6540	**1178.3001**	11	A	4	**488.5644**	244.7862
654.2118	**1307.4156**	12	E	3	**417.4856**	209.2468
710.7915	**1420.5751**	13	I	2	**288.3701**	144.6890
		14	R	1	175.2107	88.1093

Comparing Proteins



Filtering criteria listed at top; proteins that match the criteria listed below.

Experimental Annotations



- Standards-based annotation of experiments
- Data/experiment exchange format
- See tutorial on http://cpas.fhcrc.org

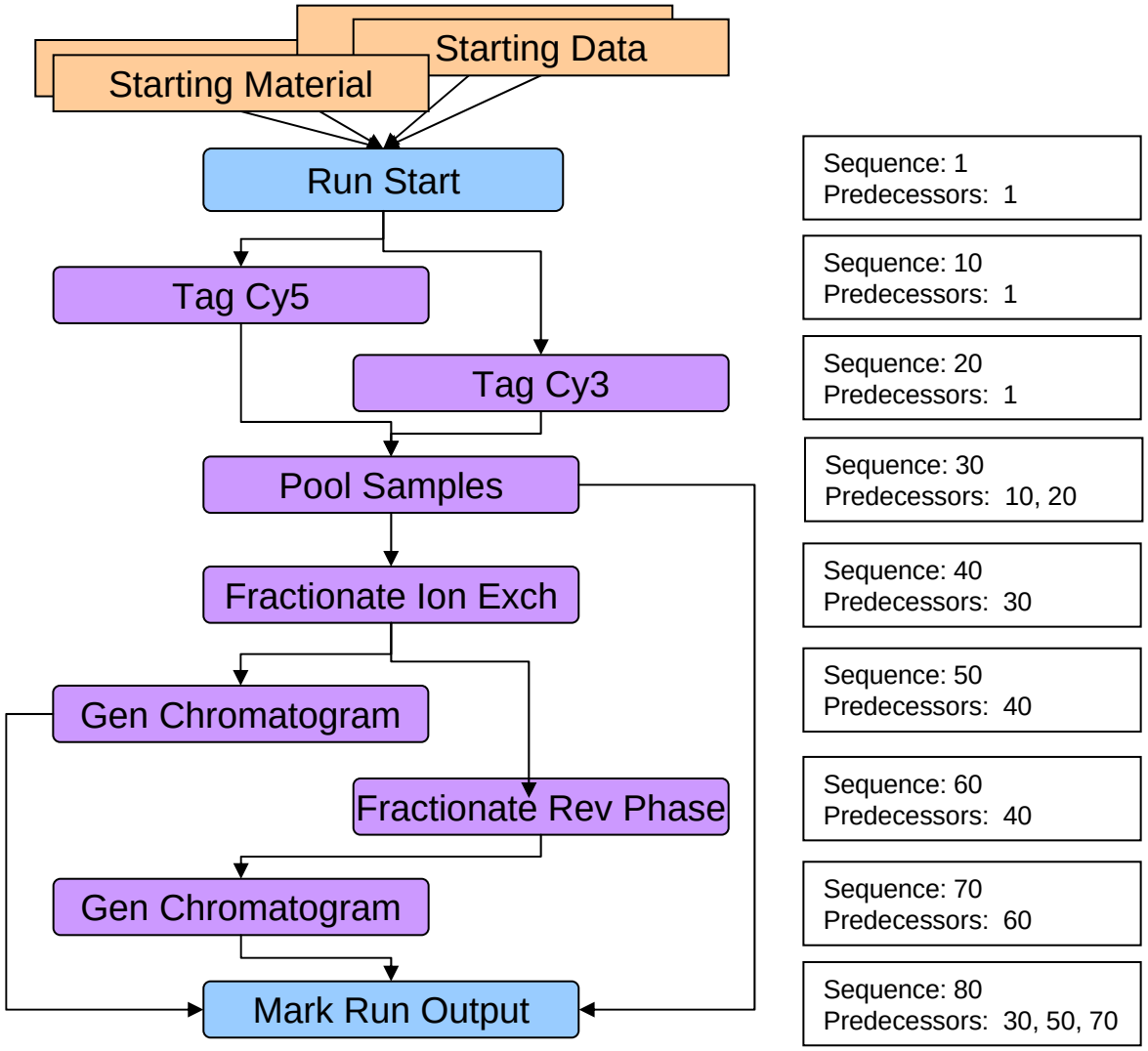
Experimental Annotations: Goals

- Dumping gigabytes of MS/MS results into a database is not enough
- Must have a framework for describing and querying experimental data in scientifically interesting ways:
  - "Show me all runs performed on Chodosh mouse model plasma samples"
  - "Across multiple mouse models, show me all differentially regulated proteins grouped by cancer-type"
  - "Show me experiments that used the glyco-capture method where protein X was found"
- Needs to separate structure:
  - inputs, protocol steps, outputs, relationships
- …from vocabulary:
  - properties/types specified by scientist or standardized ontologies
- Requires flexibility
  - Database schema, file formats, and tools must support constantly changing protocols, terms, properties, and ontologies

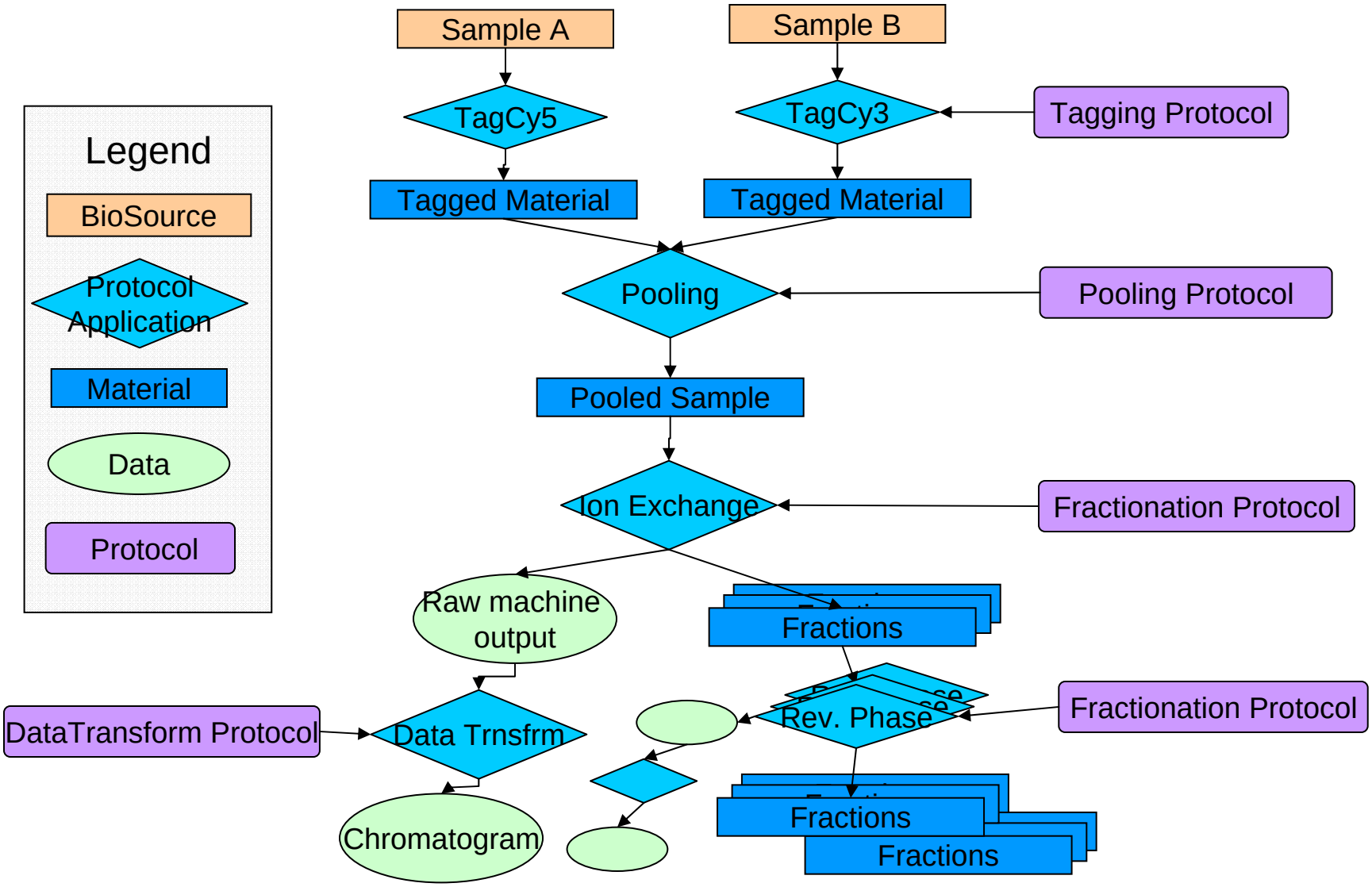
Solution Components

- Experiment Archive File: *myexperiment.xar*
  - All data files and manifest zipped together
- Manifest file: *myexperiment.xar.xml*
  - XML doc adhering to an extensible XML Schema
  - Follows the base object structure of FuGE-OM
- Database schema to store experiment info
- Data pipeline: UI for collecting annotations and initiating server upload and processing
- Web-based query interface over database (soon)

Example: Protocol Definition



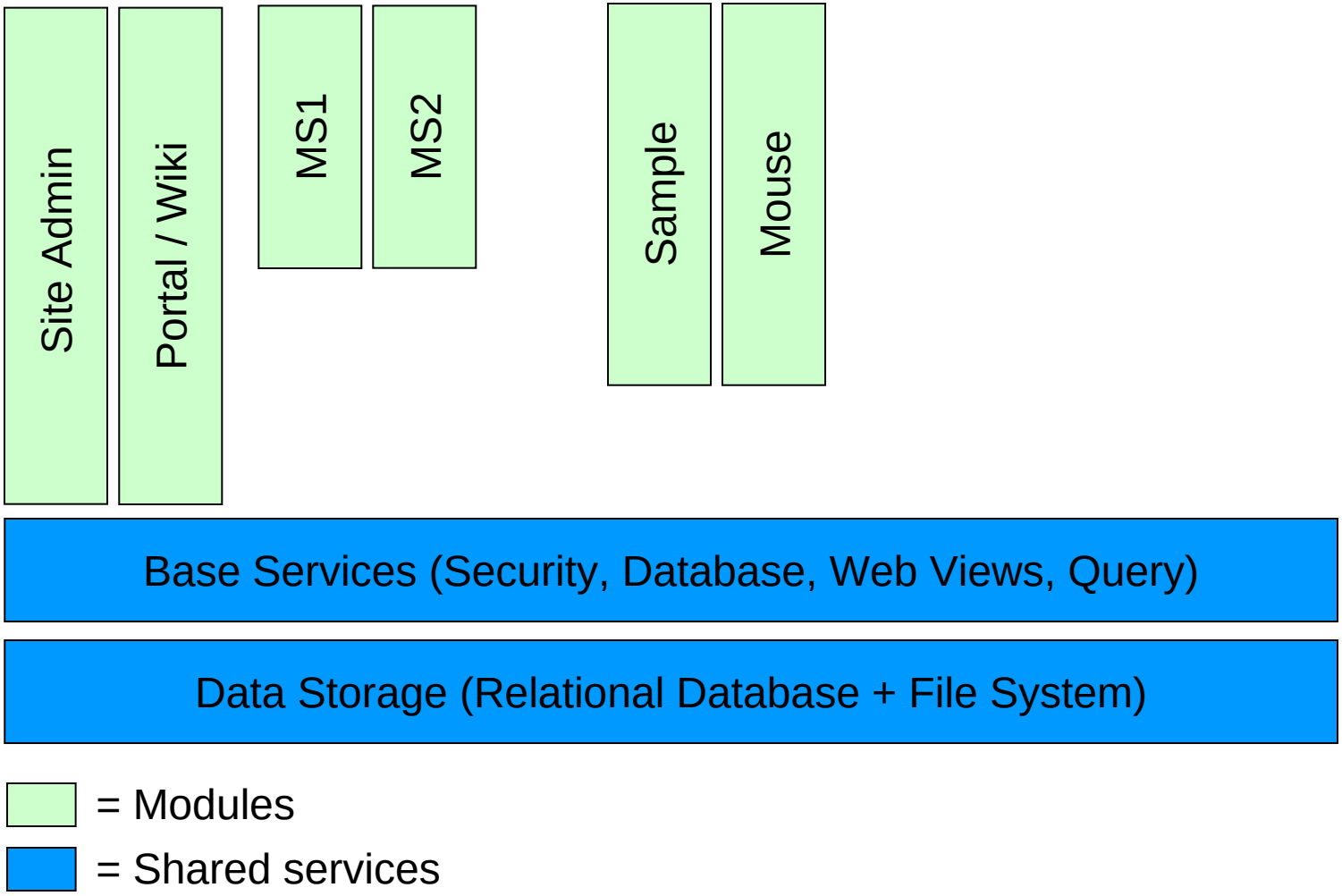
Example: Experiment Run



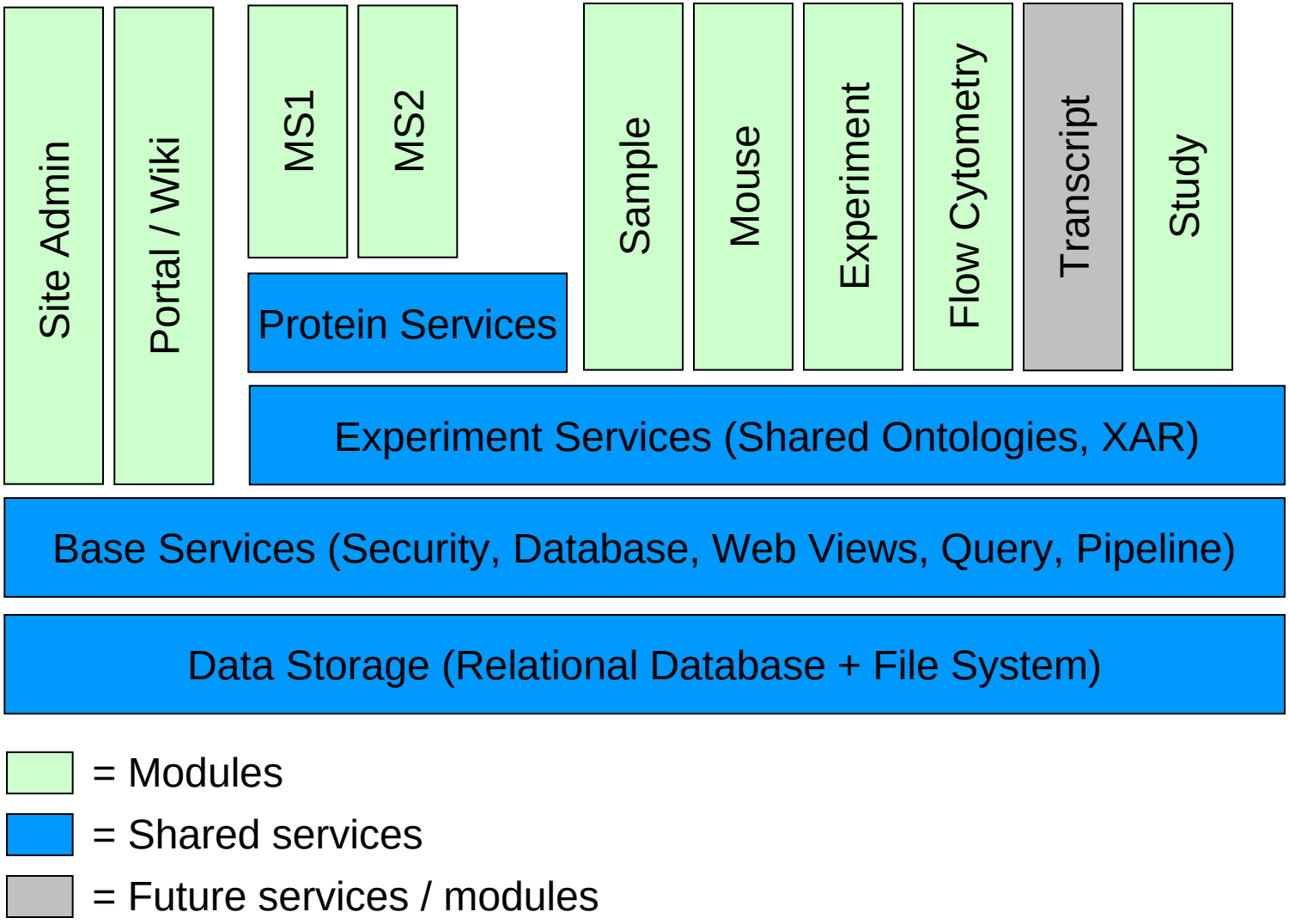
Protein Services

- CPAS links MS/MS results to database of protein sequences & annotations
  - Protein sequences are loaded from both FASTA files and annotated protein databases (e.g., UniProt)
  - Each sequence is stored once per organism and given a unique SeqID
  - All identifiers, descriptions, annotations, and references from all sources are linked to corresponding SeqID
  - Schema supports addition of new types of identifiers and annotations
- This provides ability to:
  - Display and link to biologically relevant protein information
  - Compare results searched against different FASTA files (IPI vs. NCBI)
  - Generate from results charts summarizing GO metabolic function, cellular location, and molecular function
  - Link new annotations to old results & regenerate FASTA files needed for re-analysis

CPAS Architecture (2004)



Beyond CPAS (2006)



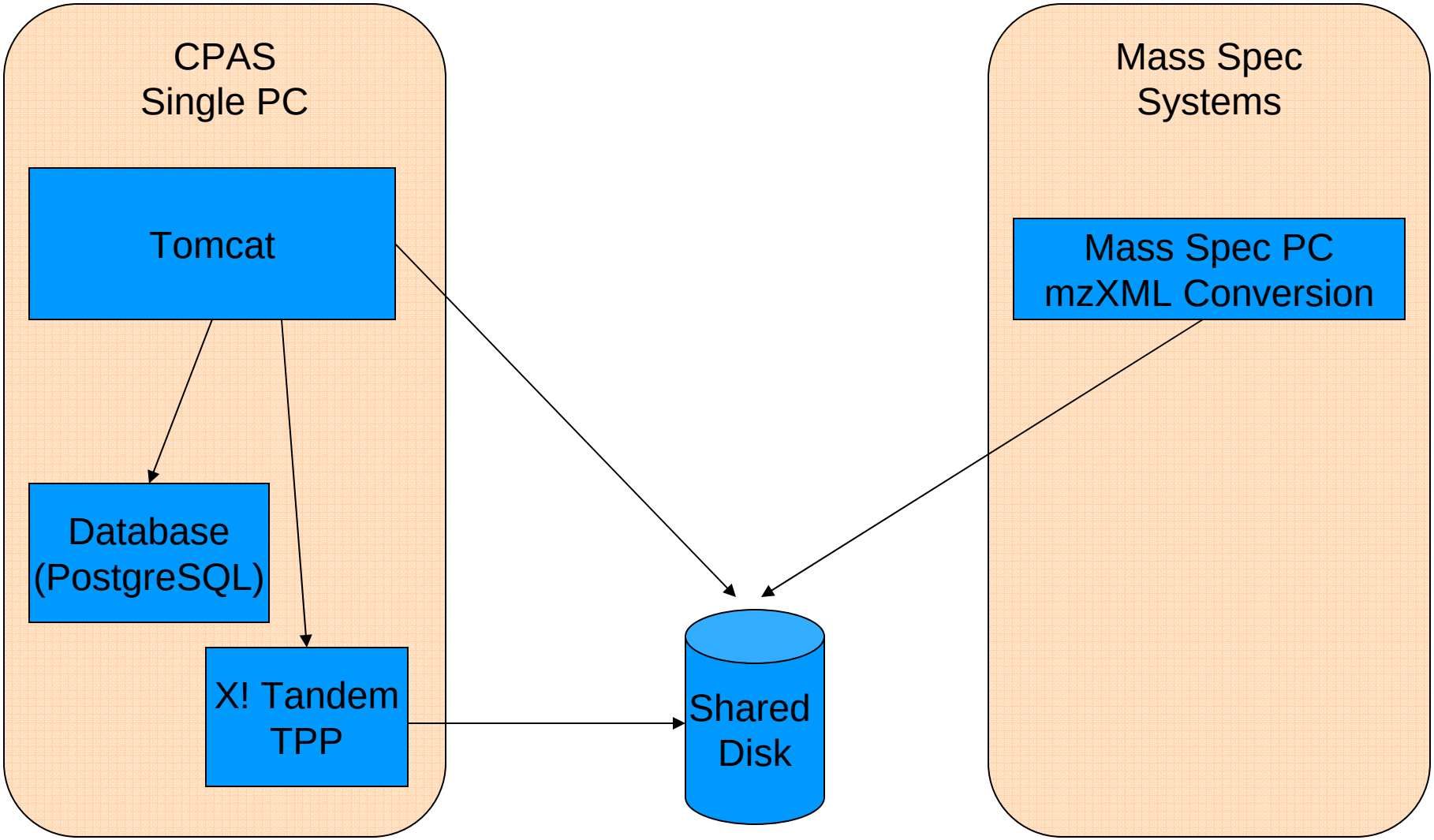
System Components

- Java web application
  - Runs on Apache Tomcat web server
  - Compatible with Windows, Linux, Solaris, Mac, et al
  - Incorporates open-source libraries
- Relational database server
  - PostgreSQL: open-source, all common operating systems
  - Microsoft SQL Server: commercial product, Windows only
  - Abstraction layer allows other database servers in future
- Network file storage: data archive
- Analysis pipeline: conversion, search, processing
- Open file formats: mzXML, pepXML, protXML, XAR

Setting Up CPAS

- Windows Installation
  - Graphical setup and configuration of “mini” MS/MS analysis system on a Windows PC:
    - CPAS application
    - Java Runtime Environment
    - Apache Tomcat
    - PostgreSQL
    - X! Tandem with multiple scoring algorithms
    - TPP components: PeptideProphet™, ProteinProphet™ , XPRESS, PepXML translators
  - Suitable for personal use, low throughput situations
- Linux Installation
  - Straightforward “manual” install of above components

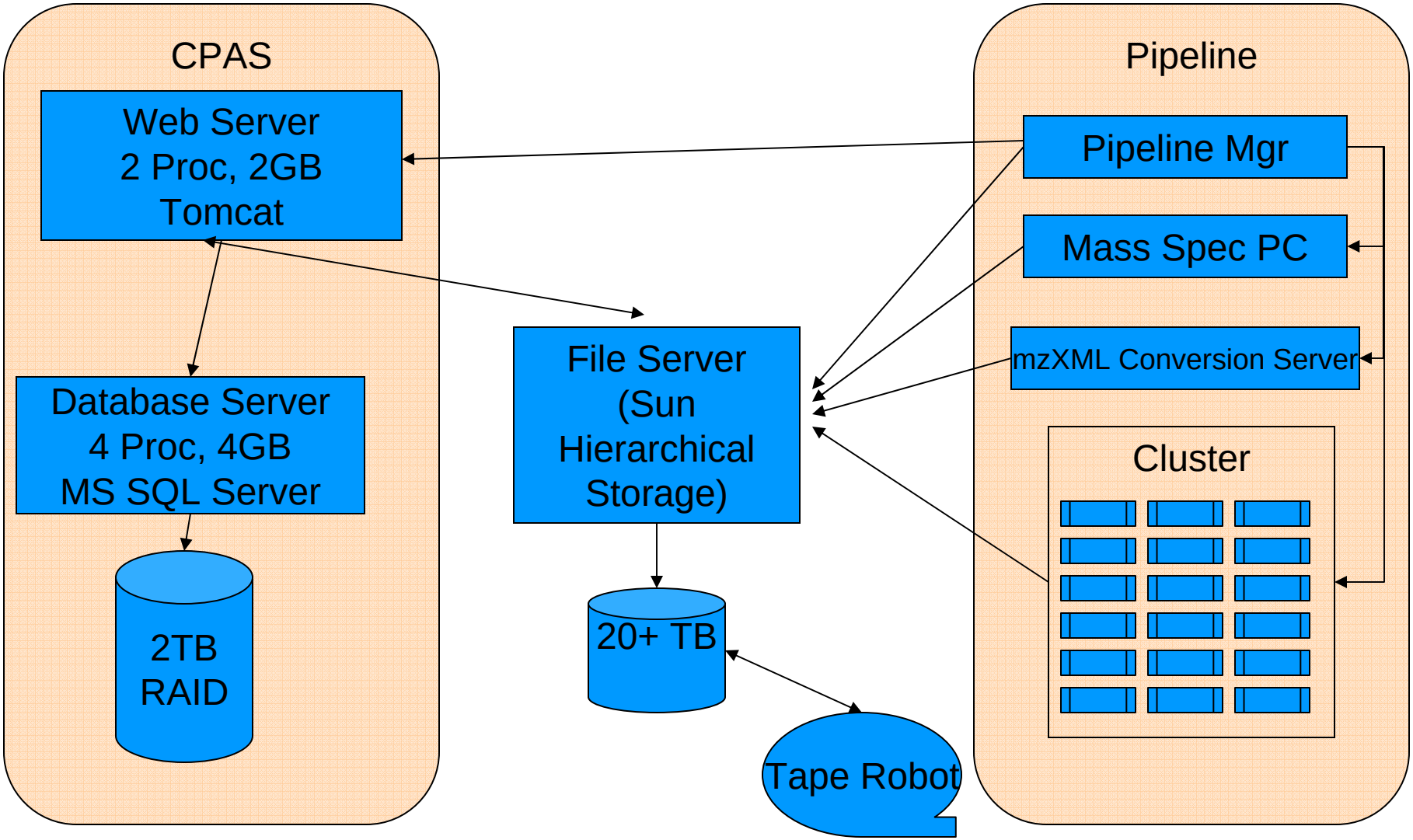
“Mini” Installation



External Pipeline

- Most proteomics facilities require more advanced setup
  - Network file system
  - Add RAW → mzXML conversion server(s)
  - Replace X! Tandem with Mascot, SEQUEST, etc.
  - Run searches and other processing on multi-node cluster
  - Additional pre- and post-search processing steps
- CPAS supports these setups
  - Configured as cron jobs & perl scripts that communicate with CPAS via log files and wget
  - FHCRC scripts are available as an example

FHCRC Installation



CPAS Pipeline Interface

- Web UI that initiates, controls, and monitors MS/MS processing
- Administrator configures pipeline
  - Pipeline root: path to RAW/mzXML file storage
  - FASTA root: path to sequence files
  - Default search parameters
- User starts MS/MS search
  - Clicks “Process and Upload Data”
  - Browses the hierarchy and selects mzXML file to process
  - Selects (or creates a new) protocol that specifies FASTA file, search & TPP parameters
  - Clicks “Search”
- CPAS then initiates and controls the data processing steps
  - Starts the MS/MS search
  - Runs the requested TPP post processing
  - Uploads the run, including experimental annotations
- User can monitor progress and status of all running jobs

Security

- Designed to keep sensitive, unpublished scientific data secure
- Admin can choose to require SSL for all access
- Authentication: dual scheme approach
  - Can delegate to institution's LDAP system
  - External users: invitation only
    - Users choose their own passwords
    - Hash of password is stored in database and used for authentication
- Authorization: Users must be granted explicit permissions
  - All data stored in folder hierarchy managed by the database
  - Users are added to groups
  - Groups are granted permission to folder or hierarchy
  - Authorized only if user belongs to group with required permissions
- Folders can be made "public" (no authentication required)

Administration UI

- Customize site
  - Organization & system names, logos, icons, support links
  - LDAP & database configuration, SSL
- Manage users
  - Add, delete, update profile, reset password, change email, history
- Manage groups and permissions
  - Create, delete groups
  - Manage group membership
  - Assign permissions
- Manage folders
  - Create, rename, move, delete
- Pipeline
  - Configure cluster pipeline
  - Select network file system root associated with each folder
  - Monitor in-progress jobs
- MS/MS
  - View statistics about runs, FASTA files
  - Purge deleted runs

CPAS Summary

- Easy way to install MS/MS pipeline and analysis system
- Ships and integrates with X! Tandem search engine & some TPP tools
- Compatible with SEQUEST & Mascot as well
- Allows storing, analyzing, mining, publishing, and exporting MS/MS results
- Supports high-throughput facilities and large collaborations
- Ties results to experimental & protein annotations
- Extensible – add your own modules

Resources

CPAS distribution and support site	http://cpas.fhcrc.org
FHCRC CPL	http://proteomics.fhcrc.org
LabKey Software	http://www.labkey.com

CPAS Paper

Rauch A, Bellew M, Eng J, et al. Computational Proteomics Analysis System (CPAS): An Extensible, Open-source Analytic System for Evaluating and Publishing Proteomic Data and High throughput Biological Experiments. J Proteome Res 2006;5(1):112-121.

Acknowledgements

- National Cancer Institute
- Canary Foundation
- ISB: TPP, mzXML, pepXML, protXML
- Ron Beavis & The GPM: X! Tandem
- Many other open-source developers

Questions?