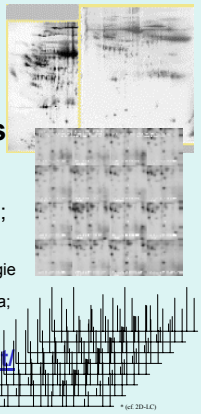
The Open2Dprot Proteomics Project for n-Dimensional Protein Expression Data Analysis



Peter F. Lemkin[1]; Jai Evans[2]; Rephael Wenger[3]; Djamel Medjahed[4]; Greg C. Thornwall[5]

[1]National Cancer Institute, Frederick, MD; [2]Carnegie Mellon Univ., Pittsburgh, PA; [3]The Ohio State Univ., Columbus, OH; [4]Univ. Algiers, Algeria; [5]SAIC-Frederick, Frederick, MD

http://open2dprot.sourceforge.net/

HUPO USA - Boston, March 12-15, 2006

Revised 3-02-2006

Abstract

There is a need for integrated proteomics expression databases and bioinformatic tools that help perform exploratory data analysis and data mining in the context of the large number of high-quality characterization, annotation, pathway and functional databases increasingly available on the Internet. Some of the biological problems addressed by these types of bioinformatic tools include aid in the detection and better understanding of post-translational modifications; helping in the discovery of biomarkers for diagnosis and monitoring of disease, detecting toxicity, and developing new drugs; analysis of coordinated expression of sets of proteins; and pathway elucidation. The Open2Dprot project is a community effort to create a fully open-source n-dimensional protein expression data analysis system that can be freely downloaded and used for data mining protein expression profiles across sets of n-dimensional data from research experiments (2D gels, 2D LC-MS, protein microarrays, n-dimensional LC-MS*MS*..., etc). The focus of Open2Dprot is to provide an integrated set of open source software tools for n-D database analysis that is hosted on the SourceForge.net repository. A pipeline control program called Open2Dprot analyzes, schedules, and runs the pipeline modules required to pre-process and create a Composite Sample Database used in the data mining.

Open2Dprot is being expanded to handle data from other protein separation methods. It uses the open source methodology modeled after our MAExplorer DNA microarray analysis software. The Open2Dprot goals and software development plan are described on http://open2dprot.sourceforge.net/. Open2Dprot is being written in Java/R using XML and MySQL RDBMS. It is based partly on some refactored code from earlier C/Unix/X-windows 2D PAGE data-mining systems, in part on code from other open-source bioinformatic software projects (such as Bioconductor), and Java and R languages using code from MAExplorer, Flicker, and GELLAB-II. It is being extended with other 2D-proteomics analyses, mass spectrometry, protein microarray, and related proteomics software codes as well as developer efforts donated by the research community. It uses XML interchange formats and a SQL/schema modeled after the Protein Standards Initiative (PSI) MIAPE proteomics community data standard as then interface between stages of the analysis pipeline. This standardization allows for data sharing and alternate methods for 2D gel spot segmentation or 2D LC-MS peptide "spot" clusters, protein-arrays, spot pairing, data analysis methods, etc., could be made added. This will be critical when applying it to other types of 2D proteomics data. As pipeline components become usable, they are made available on the Web site (see 'Module list' for current status).

The Open2Dprot Project

Open2Dprot is an open-source project for the development of n-dimensional proteomics exploratory data analysis bioinformatic tools.

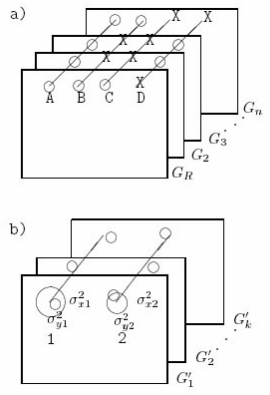
The tools can be used for analyzing quantified protein expression data across multiple n-D samples from research experiments.

The tools could be adapted for use with a variety of quantified 2-D or n-dimensional protein separation sources of expression data.

Proteomic Separation Methods

- **2D-PAGE** (P. O'Farrell, 1975) pIe vs Mm (mass), 2D-gels
  2D LC-MS retention-times vs m/z (mass)
  2D IPG-MS pIe vs m/z (mass)
  2D LC-LC pIe vs RP-HPLC
  n-D (e.g., LC-MS*MS*MS …)
  Protein arrays (analytes vs antibodies**)**
- All share a common paradigm: proteins separated by orthogonal features
- Some of these methods are semi-quantitative
- Data represented as protein expression profiles lends itself to exploratory data analysis
- Open2Dprot could be used as part of a broader set of integrated tools

Composite Samples Database (CSD) Paradigm



Proteomic composite samples database (CSD) consisting of a set of n samples G_1, G_2, …,G_n with representative sample G_r=G_1

Expression profiles A,B,C, ... O = present, X = missing

A canonical sample database is a statistical representation of the CSD spot geometry and quantification that could be used for data mining

Lemkin & Lester *Clinical Chemistry*, 1982

n-Dimensional Protein Expression Analysis

PRE-PROCESSING PIPELINE:
Samples are accessioned and processed in a data-reduction pipeline to construct a Composite Samples Database CSD merging paired "spot" lists from the set of samples

Interactive and batch processing of samples

↓

DATABASE:
CSD is created & maintained in RDBMS could be shared between groups
XML files – data interchange
High-speed caches used for data mining

↑

DATA MINING:
The CSD data is explored using exploratory data analysis techniques: statistics, clustering, classification, direct-manipulation graphics and reports, etc. with access to Internet proteomic / genomic / PubMed / function / pathway databases

Interactive and batch processing of samples

7

Pre-Processing Pipeline Steps

- Accession samples (images or other data) and experiment data into database
- Quantify 'spots' from sample images, 2D LC-MS peptide peak clusters, protein arrays, or other data
- Pair spots between samples and reference samples (if needed)
- Construct Composite Samples Database (CSD) for all sets of paired samples

8

Data-Mining Analyses Being Developed for the Composite Samples Database

- Manage experimental samples containing lists of proteomic "spot" data
- Protein annotation from Internet servers from proteomic, genomic, PubMed, GO, functional & pathway DBs – used for clustering and other analyses
- Manage multiple samples as named condition sets of samples (e.g., replicates, time series, drug-dose, etc.
- Normalize data by a variety of within-sample and between-sample methods
- Select, find, and manage named subsets of the CSD using sets of samples, condition sets of samples, and sets of proteins for further processing
- Analyze protein expression profile data for these CSD subsets
- Cluster proteins or cluster samples based on various similarity criteria
- Classify samples by protein subsets, cross-validation, false-discovery corrections
- Data-filter protein sets by statistics, clustering, and protein subset membership
- Direct-manipulation data in graphics (scatter plots, expression profile plots, histograms, clustergrams or heat-maps, PCA, MDS, etc.), spreadsheets, samples

9

How: Data Mining the Composite Sample Database

- Many of these data-mining tools will be developed for Open2Dprot CSD data mining using Java- and R-dynamic plugins technology we had developed previously for MAExplorer.sf.net DNA microarray software

- In Open2Dprot, many of the R-plugins will use methods developed for or derived from Bioconductor (see bioconductor.org, DNA microarray analysis system written in the R language, r-project.org)

10

Why 2D-Gels Now?

- 2D-PAGE was not widely used until recently due to:
  - limitations in identifying spots differentially expressed
  - difficulty resolving and detecting specialized classes of proteins (e.g., basic proteins, membrane proteins, low abundance proteins)

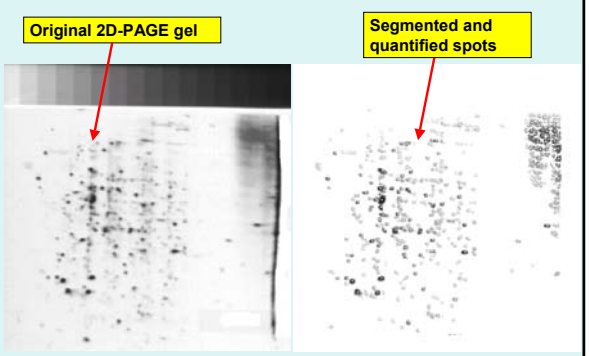
- Today, 2D-PAGE is often used as prescreening stage for mass-spectrometry to identify excised spots found in differential analysis

- Improved resolution: zoom 2D-gels, new pre-fractionation methods

- There are other protein separation techniques that could use these 2D-gel and recent DNA-microarray database analysis paradigms including 2D LC-MS and protein arrays

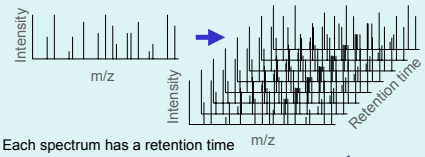
11

2D-PAGE Gel Spot Quantification Data



12

2-D LC-MS Map of Spectral Data



Each spectrum has a retention time for when is was collected

Retention times are reproducible measures of when peptides are released from the RP column

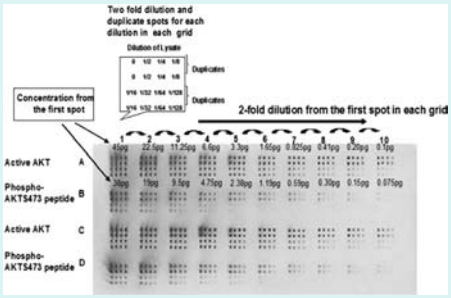
Retention time and m/z can be used as coordinates

Intensity values for each peak are mapped to a specific location in a 2-D space

* John Lewis, Geo-Centers Inc, US Army Center for Environmental Health Res., Nov, 2003 – preliminary data (with permission)

Protein Array Data

Example of sensitivity and reproducibility analysis of the reverse phase protein microarrays. From Sheehan, K. M. (2005) Mol. Cell. Proteomics 4: 346-355. With RP arrays, analytes are immobilized in solid-phase on the array.



Why Open Source?

"The basic idea behind open source is very simple: When programmers can read, redistribute, and modify the source code for a piece of software, the software evolves. People improve it, people adapt it, people fix bugs. And this can happen at a speed that, if one is used to the slow pace of conventional software development, seems astonishing.**"**

"We in the open source community have learned that this rapid evolutionary process produces better software than the traditional closed model, in which only a very few programmers can see the source and everybody else must blindly use an opaque block of bits.**"**

From the Open Source Initiative (OSI)

http://www.opensource.org/

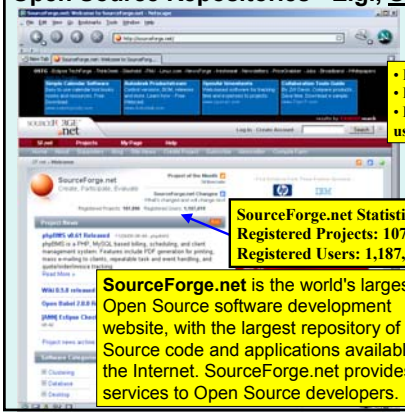
Why an Open-Source nD-Data Proteomics Effort?

- "***An open-source project can be advantageous to the community at large, since there is a far greater likelihood of progress in algorithm design in an academic style collaboration than a closed-source business model***."
- Researchers can more rapidly adapt new methods to existing software without waiting for release of commercial products
- Use contributed expertise and code of proteomics experts and bioinformaticians to help build and test open software
- Algorithms more transparent, so researchers can verify results more easily
- More opportunity to share data in standard non-proprietary formats

Why Open Source Proteomics? (continued)

- No expensive software licenses required - reduces deployment costs within large organizations and small labs
- Using proper open-source licenses can encourage adoption and collaboration between industry, academic, and government interests (e.g., Linux, FireFox, Apache, Eclipse etc.)
- Many free open-source repositories available
- Repositories offer tools to support collaboration, software development, documentation, forums, and distribution

Open Source Repositories - E.g., SourceForge.Net



Open2Dprot - Project Goals

- An international community effort to create an open-source n-D quantitative data analysis system
- A stand-alone downloadable system that can connect to DBs
- Use for data mining protein expression data sets of samples from researcher's experiments to investigate and find significant protein expression differences from multiple experimental conditions
- Will provide integrated set of software tools, analysis methods and data structures for quantitative and system biology protein expression
- Will handle protein expression data from 2D-gel, 2D LC-MS, protein arrays, and other protein separation methods

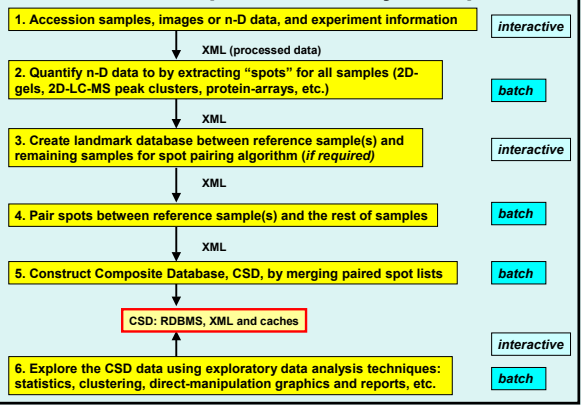
Development Plan

- Open2Dprot is being written in Java and R languages using XML (MIAPE proteomics schema) and MySQL RDBMS - modern modular open-source technologies aiding portability and extensibility
- Open2Dprot was derived from new and refactored Java code from various projects including: MAExplorer, Flicker, GELLAB-II
- Data mining will use Java- and R-plugins derived from MAExplorer and R data-mining open-source proteomics (e.g., Bioconductor) , as well as other bioinformatics data-mining software
- Will extend with other open-source analysis related proteomics software codes with additional efforts by research community

Using Open Source Resources

- Hosted and developed on SourceForge repository at **open2dprot.sourceforge.net**
- Web site discusses the Open2Dprot software development plan, and contains documentation and software distributions
- Uses the similar open-source development methodology used in our Java/R-based MAExplorer **maexplorer.sourceforge.net** DNA microarray data-mining software
- Open2Dprot could later reside as part of analysis or other reference database Web sites integrated with other proteomics tools

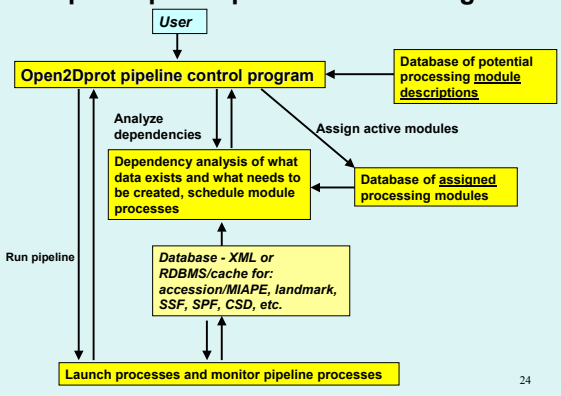
n-D Protein Expression Analysis Pipeline



Pipeline Control Program – Open2Dprot

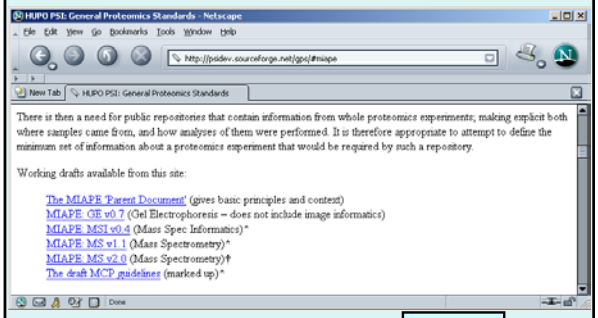
- The pre-processing is controlled by the pipeline control program "Open2Dprot"
- Modules are assigned to Processing stages
- It determines what data exists; what data needs to be created from existing data; and creates the "target" data from that dependency by future module processing
- It then schedules and runs the required dynamically assigned modules in the pipeline to create the target data. Multiple processors could be used
- This is repeated until the desired data is created

Open2Dprot Pipeline Control Program



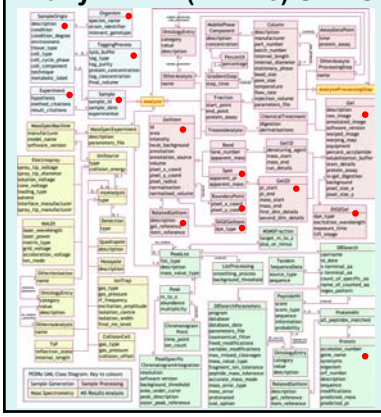
MIAPE – Minimal Information About a Proteomics Experiment

psidev.sourceforge.net/gps/#miape



As of 2-14-2006

Early MIAPE (PEDRo) UML Schema n-D Data



- Classes that could be used with 2D-gels

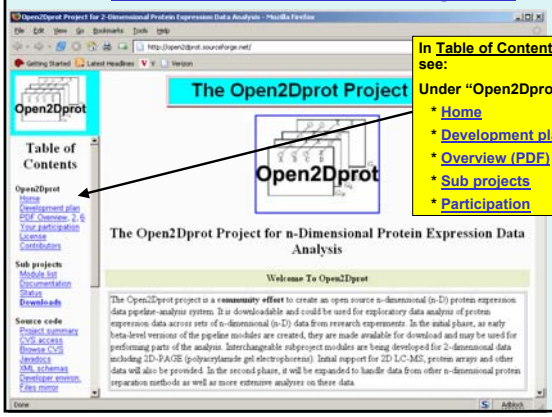
Additional fields / classes are needed for Open2Dprot

in Taylor et.al., *Nature Biotechnology*, March 2003.

PEDRo has been renamed MIAPE "Minimal Information About a Proteomics Experiment" (Oct. 2003, HUPO-II) by EMBL-EBI

See psidev.sf.net

Home: http://open2dprot.sourceforge.net/



In Table of Contents, see:

Under "Open2Dprot"

* Home
* Development plan
* Overview (PDF)
* Sub projects
* Participation

Open2Dprot Pipeline Subprojects - Status

Subproject Home	Download	Documentation	Overview (PDF)	PDF documents	Version	Revision history	Status	Pipeline step
Open2Dprot	(see below)	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot overall design	[Overall design]
Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot pre-alpha program	[scheduled]
Accession	Accession	Accession	Accession	Accession	Accession	Accession	Accession Beta	[1]
Seq2Pael	Seq2Pael	Seq2Pael	Seq2Pael	Seq2Pael	Seq2Pael	Seq2Pael	Seq2Pael Beta	[2]
Landmark	Landmark	Landmark	Landmark	Landmark	Landmark	Landmark	Landmark Beta	[3]
AutoLandmark	AutoLandmark	AutoLandmark	AutoLandmark	AutoLandmark	AutoLandmark	AutoLandmark	AutoLandmark pre-alpha	[3]
CmpSpots	CmpSpots	CmpSpots	CmpSpots	CmpSpots	CmpSpots	CmpSpots	CmpSpots Beta	[4]
BuildCSD	BuildCSD	BuildCSD	BuildCSD	BuildCSD	BuildCSD	BuildCSD	BuildCSD pre-alpha	[5]
CSDminer	CSDminer	CSDminer	CSDminer	CSDminer	CSDminer	CSDminer	CSDminer design prototype	[6]
O2Plib	O2Plib.jar	O2Plib	O2Plib	O2Plib	O2Plib	O2Plib	O2Plib Beta	--common--

Additional alternative modules are being developed for all pipeline stages

01-12-2006

Contributed Associated or Related Projects

We added some additional non-pipeline open source projects that may use similar data or common software modules. They may be useful for performing other types of analysis on data used by Open2Dprot or provide other types of analyses.

Contributed Project Home	Download	Documentation	Overview (PDF)	PDF documents	Version	Revision history	Status
Flicker	Flicker	Flicker	Flicker	Flicker	Flicker	Flicker	Flicker
MAExplorer	MAExplorer	MAExplorer	MAExplorer	MAExplorer	MAExplorer	MAExplorer	MAExplorer
ProtPlot	Protplot	TMAP (ProtPlot)	ProtPlot	ProtPlot	ProtPlot	ProtPlot	...

01-12-2006

Summary of Open2Dprot

- Open2Dprot is a fully open-source n-D proteomics data-mining project for a variety of proteomic expression data sources and is being developed at http://open2dprot.sourceforge.net/
- It has a flexible pipeline-modules design using XML data interchange and /RDBMS-caches and portable Java and R using existing code where possible
- As parts of the project pipeline become usable, they are being released as stand-alone programs