## Open2Dprot

## The Open N-Dimensional

## Proteomics Project

Peter Lemkin
Lab Exper. & Computational Biology
NCI-Frederick, MD

**http://open2dprot.sourceforge.net/**

PROTIG Seminar, NIH, March 17, 2005

---

## Overview

- **1. The Open2Dprot project**
- **Definition and project goals**
- Open source
- Development plan
- Analysis pipeline

- 2. Examples of some of the pipeline modules

- 3. Open2Dprot data mining model. E.g., MAExplorer, Java- and R-plugins

2

---

## 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.
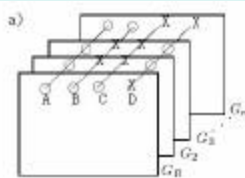
3

---

## 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

- Could be used for data mining protein expression across sets of samples from researcher's experiments to investigate and find significant protein expression from multiple experiments

- 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
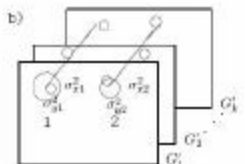
4

---

## Composite Samples Database (CSD) Paradigm



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

Expression profiles A,B,C, ...
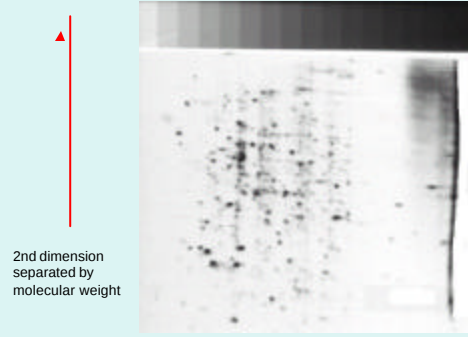
A canonical sample database is a statistical representation of the CSD spot geometry and quantification that could be used for data mining

in Lemkin *et al.*, *Computers Biomedical Research*, 1981

5

---

## 2D Polyacrylamide Gel Electrophoresis Separation
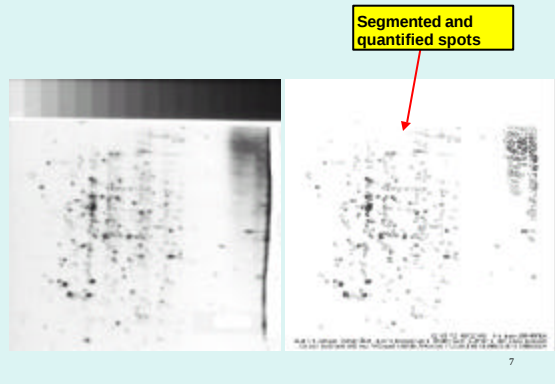


2nd dimension separated by molecular weight

1st dimension separated by isoelectric point (pH)

6

## 2D-Gel Spot Segmentation and Quantification

Segmented and quantified spots



7
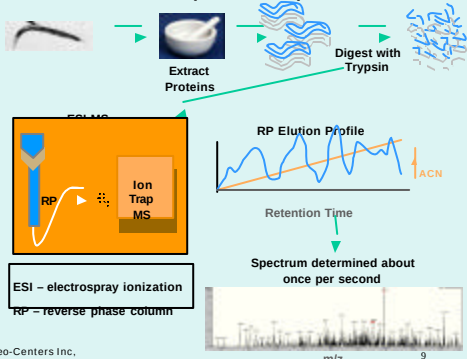
## Why 2D-Gels Now?

- 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 <u>prescreening stage</u> for mass spectrometry to identify spots found in differential analysis

- <u>Improved resolution</u>: zoom 2D-gels, new pre-fractionation methods

- There are <u>other protein separation techniques</u> that could use these 2D-gel and recent DNA-microarray database analysis paradigms
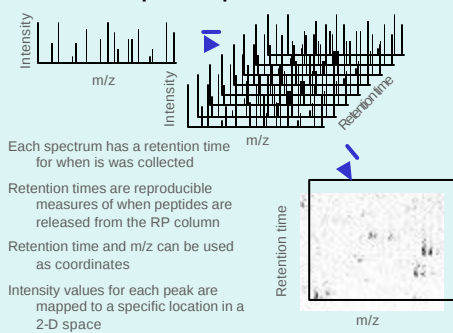
8

## LC-MS work flow for protein quantification
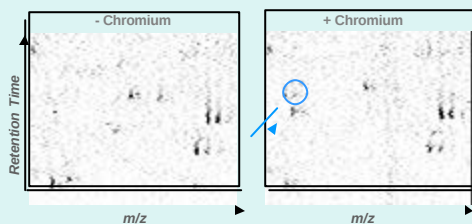


**Extract Proteins**

**Digest with Trypsin**

ESI-MS

**RP Elution Profile**

ACN

RP

**Ion Trap MS**

**Retention Time**

**Spectrum determined about once per second**

**ESI – electrospray ionization**

**RP – reverse phase column**

m/z

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

9

## 2-D Map of Spectral Data

Intensity

m/z

Intensity

m/z

Retention time

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

Retention time

m/z

10

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

## Screening MS only data for peptide biomarkers

- Chromium

+ Chromium

Retention Time

m/z

m/z
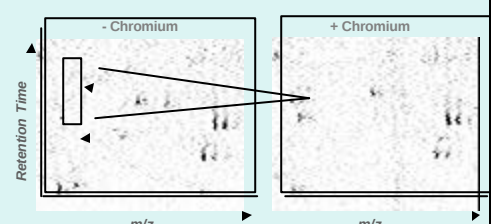
Peak m/z and retention time are reproducible
- Variability - measurement error and chromatographic effects

Intensity more susceptible to instrumental variation
- Acceptable if no changes to MS (including getting dirty)
- Ideally, samples run consecutively

11

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

## Screening MS only data for peptide biomarkers (2)

- Chromium

+ Chromium

Retention Time

m/z

m/z

Use centroid data to reduce number of points

Scan each point against a window in an other sample – store point if n-fold greater than any point in other sample window

Identify cluster with multiple points that are likely to form a peak

Use m/z and retention time to perform target MS/MS analysis to identify peptide

12

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

## Composite Samples Database (CSD) Paradigm



**Proteomic composite samples database (CSD) consisting of a set of n samples $G_1, G_2, …, G_n$ with <u>representative</u> <u>sample</u> $G_r = G_1$**

**Expression profiles A,B,C, ...**

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

in Lemkin *et al.*, *Computers Biomedical Research,* 1981     13

---

## Overview

- **1. The Open2Dprot project**
- Definition and project goals
- **Open source**
- Development plan
- Analysis pipeline

- 2. Examples of some of the pipeline modules

- 3. Open2Dprot data mining model. E.g., MAExplorer, Java- and R-plugins

14

---

## Why Open Source?

"The <u>basic idea behind open source</u> 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 <u>rapid evolutionary process produces better software</u> 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/**     15

---

## Open Source Repositories - E.g., <u>SourceForge.Net</u>



**Free code Repositories Developer, collaborator, ... environments**

**SourceForge.net Statistics Registered Projects: 97,306 Registered Users: 1,034,516**

**SourceForge.net** is the world's largest Open Source software development website, with the largest repository of Open Source code and applications available on the Internet. SourceForge.net provides free services to Open Source developers.

03-16-2005

16

---

## Using Open Source Resources

- Initially, <u>hosted and developed</u> on SourceForge.Net repository at **open2dprot.sourceforge.net**

- This Web site discusses the current Open2Dprot <u>software development plan</u>

- Use the same <u>open-source development methodology</u> used in our Java/R-based MAExplorer **maexplorer.sourceforge.net** DNA microarray data-mining software

- Open2Dprot could could later be used or integrated with other proteome exploratory or database tools

17

---

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



**In <u>Table of Contents</u>, see: under "Open2Dprot"**
**\* Home**
**\* Development plan**
**\* Overview (PDF)**
**\* Subprojects**

## Overview

- **1. The Open2Dprot project**
- Definition and project goals
- Open source
- **Development plan**
- Analysis pipeline

- 2. Examples of some of the pipeline modules

- 3. Open2Dprot data mining model. E.g., MAExplorer, Java- and R-plugins

19

---

## Development Plan

- Open2Dprot is being written in Java and R languages using XML and MySQL RDBMS - modern modular open-source technologies aiding portability and extensibility

- Initially Open2Dprot is being derived from refactored code to Java / R / XML / MySQL-RDBMS paradigm:

  **a**) parts of NCI GELLAB-II system - the C-language / Unix / X-windows (1993) **www.lecb.ncifcrf.gov/gellab**,

  **b**) Java and R plugins from MAExplorer **maexplorer.sourceforge.net**, R data-mining software **www.r-project.org** such as Bioconductor **bioconductor.org**,

  **c**) from other open source proteomics and bioinformatics projects.
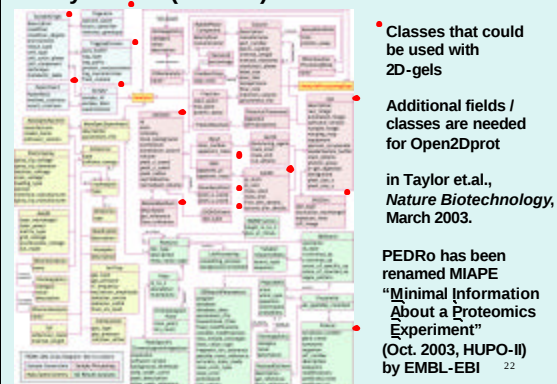
20

---

## Development Plan (cont.)

- Extend with other 2D-gel, LC-MS$^N$, protein array and related open-source proteomics software codes

- Work with proteomics standardization groups (MIAPE - formerly PEDRo, PSI, HUPO, and others) to develop and use a standard database schema

21

---

## 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**
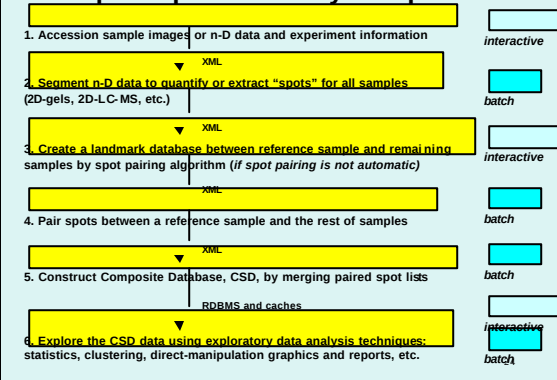
22

---

## Overview

- **1. The Open2Dprot project**
- Definition and project goals
- Open source
- Development plan
- **Analysis pipeline**

- 2. Examples of some of the pipeline modules

- 3. Open2Dprot data mining model. E.g., MAExplorer, Java- and R-plugins

23

---

## Open2Dprot n-D Analysis Pipeline

1. Accession sample images or n-D data and experiment information — *interactive*

XML

2. Segment n-D data to quantify or extract "spots" for all samples (2D-gels, 2D-LC-MS, etc.) — *batch*

XML

3. Create a landmark database between reference sample and remaining samples by spot pairing algorithm (*if spot pairing is not automatic)* — *interactive*

XML

4. Pair spots between a reference sample and the rest of samples — *batch*

XML

5. Construct Composite Database, CSD, by merging paired spot lists — *batch*

RDBMS and caches

6. Explore the CSD data using exploratory data analysis techniques: statistics, clustering, direct-manipulation graphics and reports, etc. — *interactive / batch*

---

4

## Initial Open n-D Data-Mining Tools

- **Accession n-D sample images or n-D data and experiment data**
- **Quantify 'spots' from sample images or peptide clusters**
- **Pair spots between samples and a reference sample**
- **Construct composite sample database for exploratory data analysis**
- **Manage subsets of proteins in the database**
- **Manage replicate samples and condition sets of samples**
- **Analyze expression profiles for multiple conditions**
- **Data-filter protein sets by statistics, clustering, set membership**
- **Direct-manipulation of data in graphics, spreadsheets**
- **Integrate R language statistical, clustering, classifiers, class prediction, and other methods**
- **Integrate access to Internet proteomic/genomic/function data servers for user-specified protein sets**

25

---

## Status: Open2Dprot Pipeline Subproject Modules



26

03-02-2005

---

## Associated Projects

We had 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 alternate types of analysis.



03-02-2005

27

---

## Overview

- 1. The Open2Dprot project
- Definition and project goals
- Open source
- Development plan
- Analysis pipeline

- **2. Examples of some of the pipeline modules**

- 3. Open2Dprot data mining model. E.g., MAExplorer, Java- and R-plugins

28

---

## Top-level Open2Dprot Pipeline Control Program



Assign or edit pipeline module program
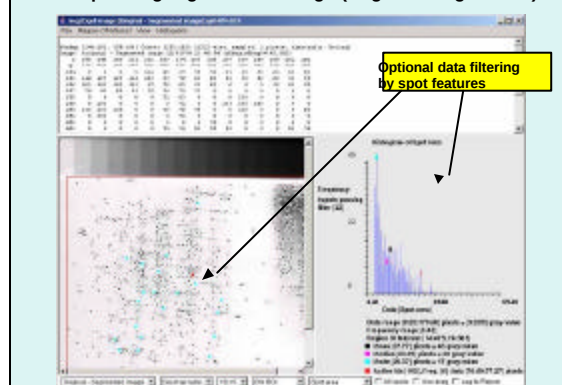
---

## Example: Seg2Dgel Image Viewer - segmented image



View quantified spot data area for selected pixel

---

5
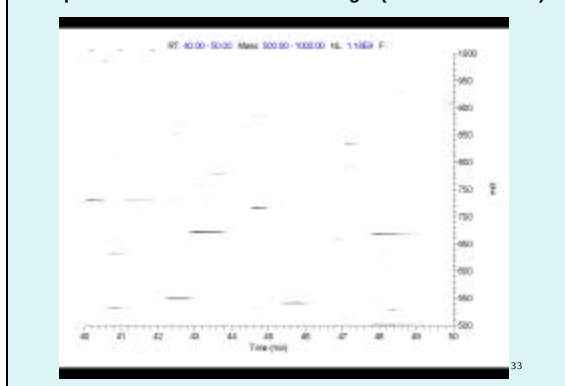
**Example: Seg2Dgel Image Viewer - background image in OD**



Window data area for selected pixel - OD
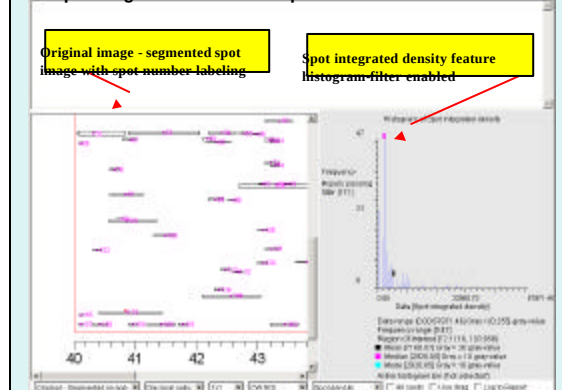
**Example: Seg2Dgel rest of image (original - segmented)**



Optional data filtering by spot features
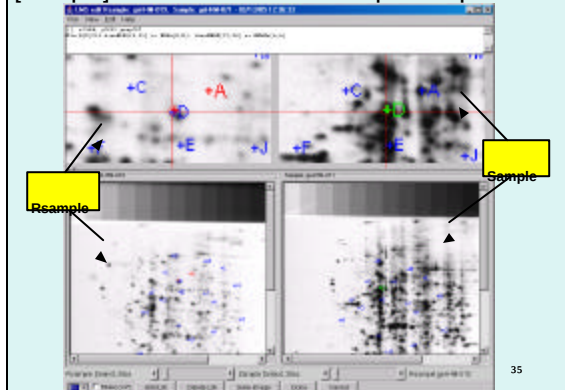
**Example: 2D LC-MS low resolution image (Veenstra/Conrads)**
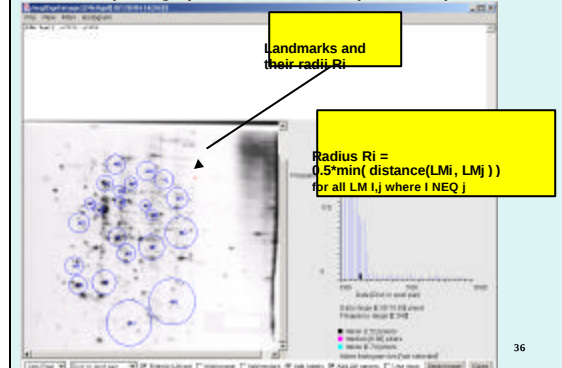


**Example: Segmentation - Semi-quantification of 2D LC-MS data**



Original image - segmented spot image with spot number labeling

Spot integrated density feature histogram-filter enabled
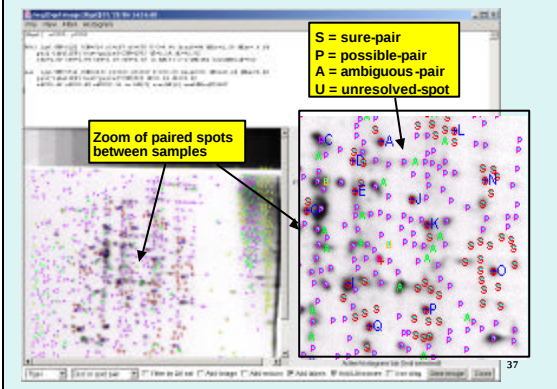
**[Pre-alpha]: Landmark – Create/Edit Rsample & Sample LMS DB**



Rsample

Sample

**Example: CmpSpots – landmark regions for pairing spots between Rsample & Sample**



Landmarks and their radii Ri

Radius Ri = 0.5*min( distance(LMi , LMj ) ) for all LM i,j where i NEQ j
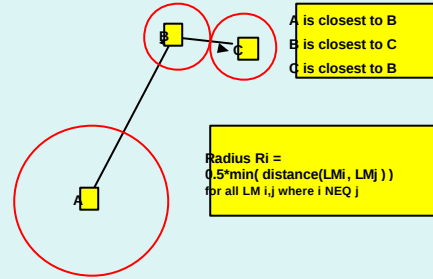
**Example: CmpSpots – paired spots between Rsample & Sample**



S = sure-pair
P = possible-pair
A = ambiguous -pair
U = unresolved-spot

Zoom of paired spots between samples

37

---

**Effective Radii of Landmarks**

Estimates of spot pairs within landmark radii are more reliable



A is closest to B
B is closest to C
C is closest to B

Radius Ri =
0.5*min( distance(LMi, LMj ) )
for all LM i,j where i NEQ j
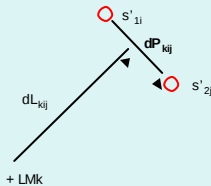
38

---

**Definition of dP - distance between spots in a pair**

The <u>distance between paired spots</u> is **dP$_{kij}$** and is computed after the two spots (s$_{1l}$,s$_{2j}$) are mapped to (s'$_{1l}$,s'$_{2j}$) in the same sample space relative to landmark k.

$$dP_{kij} = |s'_{1i} - s'_{2j}|.$$



39

---
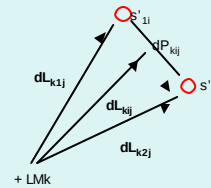
**Definition of dL - distance from a pair to it's landmark**

The <u>distance between a spot-pair and the landmark</u> is **dL$_{kij}$** and is computed after the two spots (s'$_{1l}$,s$_{2j}$) are mapped to (s'$_{1l}$,s'$_{2j}$) in the same sample space relative to landmark k.

$$dL_{k1i} = distance(LMk,s'_{1i}),$$
$$dL_{k2j} = distance(LMk,s'_{2j}),$$
$$dL_{kij} = min(dL_{k1i}, dL_{k2j}).$$



40

---
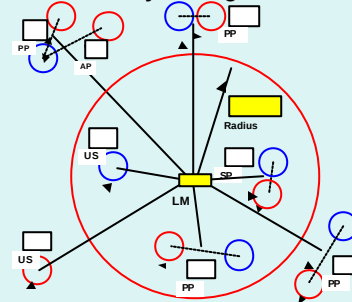
**Primary Pairing Cases**

Case
[1] US is unresolved spot (no dP),
[2] SP is dL <= R$_k$ and dP <= dTsp,
[3] PP is dL >= R$_k$ and dP <= dTpp,
[4] PP is dL <= R$_k$ and dP >= dTsp and dP <= dTpp,
[5] PP is dL >= R$_k$ and dP <= dTsp,
[6] PP is dL >= R$_k$ and dP <= dTpp. For the other spot

AP' is dL' >= R$_k$ and dP' <= dTpp and dP' >= dP

41

---

**Primary Pairing Cases**



42

---

7

## Overview

- 1. The Open2Dprot project
  - Definition and project goals
  - Open source
  - Development plan
  - Analysis pipeline

- 2. Examples of some of the pipeline modules

- **3. Open2Dprot data mining model. E.g., MAExplorer, Java- and R-plugins**
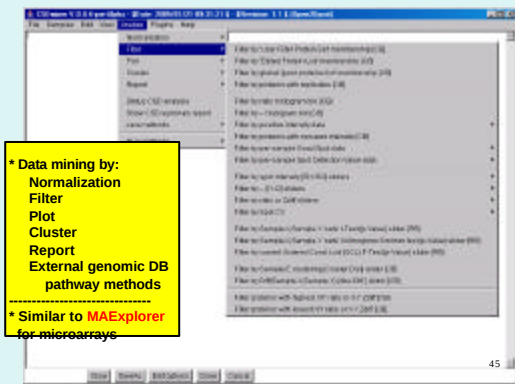
43

---

### Data-Mining Tools for Composite Sample Database

- Manage <u>subsets of proteins</u> in the database
- Manage <u>replicate samples</u> and condition sets of samples
- Analyze <u>expression profiles</u> for multiple conditions
- <u>Data-filter</u> protein sets by statistics, clustering, set membership
- <u>Direct-manipulation</u> of data in graphics, spreadsheets
- Integrate <u>R language</u> statistical, clustering, classifiers, class prediction, and other methods
- Integrate <u>access to Internet</u> proteomic/genomic/function <u>data servers</u> for user-specified protein sets

44

---

### [Future]: CSDminer – Data mining and Statistical Tool



* Data mining by:
   Normalization
   Filter
   Plot
   Cluster
   Report
   External genomic DB
      pathway methods
----------------------------------
* Similar to **MAExplorer** for microarrays

45

---

### MAExplorer  http://maexplorer.sourceforge.net/
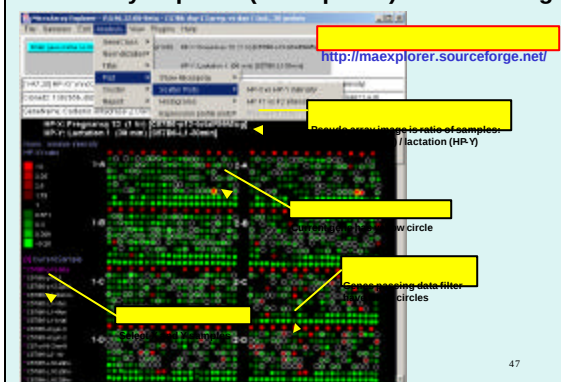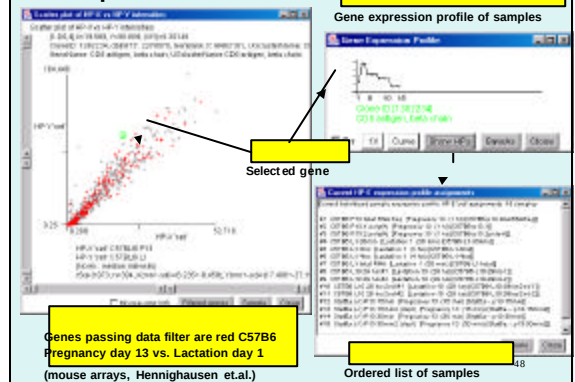


46

---

### MicroArray Explorer (MAExplorer) – Data Mining



http://maexplorer.sourceforge.net/

47

---

### MAExplorer Scatter Plot of Sample Replicates



Gene expression profile of samples

Selected gene

Genes passing data filter are red C57B6
Pregnancy day 13 vs. Lactation day 1
(mouse arrays, Hennighausen et.al.)

Ordered list of samples
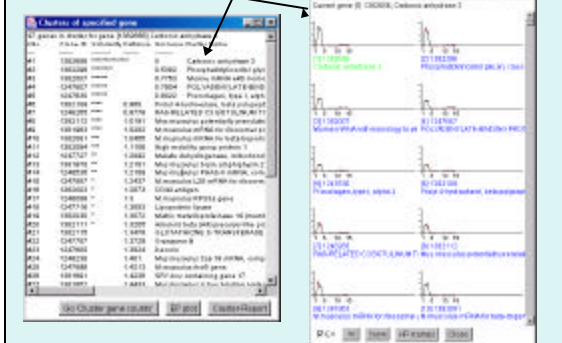
48

## MAExplorer Similarity Clustering
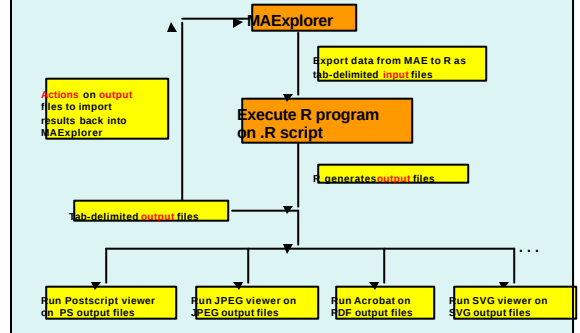
Similarity of other genes to selected gene by expression profiles

Ordered list of expression profiles



## Evaluating R Scripts on MAExplorer Data

The "R Layout" (RLO) specifies the R script, menu name, input/output files, actions, etc.

MAExplorer

Export data from MAE to R as tab-delimited input files

Actions on output files to import results back into MAExplorer

Execute R program on .R script

R generates output files

Tab-delimited output files

Run Postscript viewer on PS output files

Run JPEG viewer on JPEG output files

Run Acrobat on PDF output files

Run SVG viewer on SVG output files
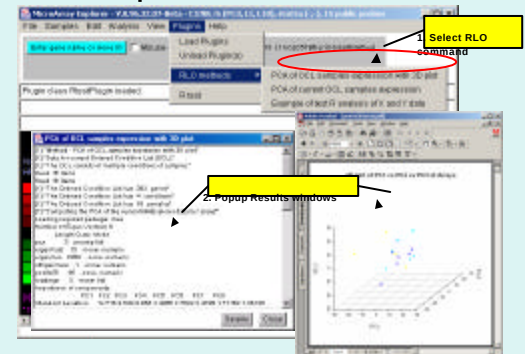


## Resources Required
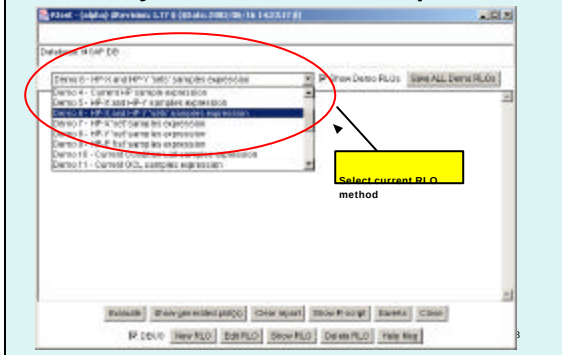
- Must install MAExplorer and the R program

- Both are open source and free
  **http://maexplorer.sourceforge.net/**
  **http://www.r-project.org/**

- MAExplorer and R run on Windows, Linux, Solaris, MacOS-X

- To view PDF generated plots, you could use Adobe Acrobat

- No other software is required

51
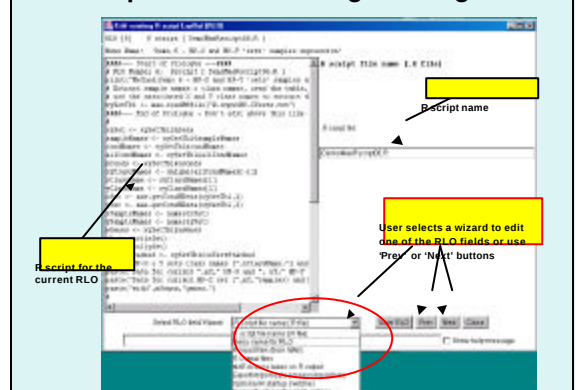
## Executing the New RLO Methods From MAExplorer RLO Methods menu

1. Select RLO command

2. Popup Results windows



## RtestPlugin: tool to Create/Edit RLO Data Analysis Methods for MAExplorer

Select current RLO method



## Multi-step Wizard for Editing/Creating RLOs

R script name

User selects a wizard to edit one of the RLO fields or use 'Prev' or 'Next' buttons

R script for the current RLO

## Summary

- 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/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

55