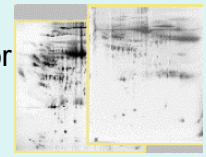
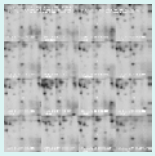
The Open2Dprot Project for n-Dimensional Protein Expression Data Analysis



Peter F. Lemkin[(1)], Gregory Thornwall[(2)]
[1]LECB, NCI-Frederick; [2]SAIC-Frederick



http://open2dprot.sourceforge.net/

HUPO-USA 2005 March 13-16, 2005, Crystal City, VA.
Open2Dprot handout
Revised: 03-03-2005

Abstract

The **Open2Dprot** project is an open-source community effort to create an n-dimensional protein expression data analysis system that can be freely downloaded and used for data mining protein expression profiles across sets of 2D data from research experiments (2D gels, 2D LC-MS, protein microarrays, n-dimensional LC-MS*MS*..., etc). The initial 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. In the future, it will be 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 in part on some refactored code from the Unix/C/X-windows version of the NCI "GELLAB-II", in part on code from other open-source bioinformatic software projects (such as Bioconductor), and Java/R languages code from MAExplorer. It will be extended with other 2D-proteomics analysis, mass spectrometry, protein microarray, and related proteomics software codes as well as developer efforts donated by the research community. It uses XML data interchange formats and a SQL/schema modeled after the developing MIAPE proteomics community data standard as the interface between stages of the analysis pipeline. This standardization allows for data sharing and alternate methods for 2D gel spot segmentation or LC-MS peptide "spot" clusters, spot pairing, data analysis methods, etc., could be made added. This will be critical when applying it to other types of 2D proteomics data.

The Open2Dprot Project

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

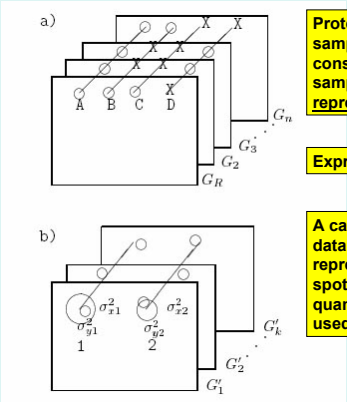
The tools can be used for analyzing quantified protein expression data across multiple n-D samples from research experiments.

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)
  n-D (e.g., LC-MS*MS*MS …)
- All share a common paradigm: proteins separated by orthogonal features
- Some methods are semi-quantitative
- Data represented as protein expression profiles lends itself to exploratory data analysis
- Open2Dprot could be used as basis for 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, ...

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

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

Why Open Source Proteomics? (continued)

- More opportunity to share data in standard non-proprietary formats
- No expensive software licenses required - reduces deployment costs within large organizations and small labs
- Using proper open-source licenses can encourage adoption and collaboration by commercial interests
- Many free open-source repositories available
- Repositories offer tools to support collaboration, software development 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
- 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, and other protein separation methods

Using Open Source Resources

- Initially, hosted and developed on SourceForge.Net repository at **open2dprot.sourceforge.net**
- This Web site discusses the current Open2Dprot software development plan
- Use the same 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 **HUPO.org** analysis Web site integrated with other tools relating to mass spectrometry, dye multiplexing, protein arrays, Internet proteomic databases, etc.

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
- Initial phase: Open2Dprot is being derived from refactored code
  a) parts of NCI GELLAB-II the C-language / Unix / X-windows
  b) from other open source proteomics and bioinformatics projects
  c) Java / R / plugins from MAExplorer and R data-mining software
- Second phase: extended with other donated 2D-gel, LC-MS^N and other analysis and related proteomics software codes with additional efforts by the research community

Development Plan (cont.)

- Work with proteomics standardization groups (MIAPE - formerly PEDRo, PSI, HUPO, and others) to develop and use a standard database schema
- Encourage research community to help expand, extend and integrate basic paradigm with other related protein separation methods and data analysis methods
- During initial phase, we especially welcome suggestions for modifying this agenda for Open2Dprot as well as core-bioinformatics developers offering to help with the project

Basic Open 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*

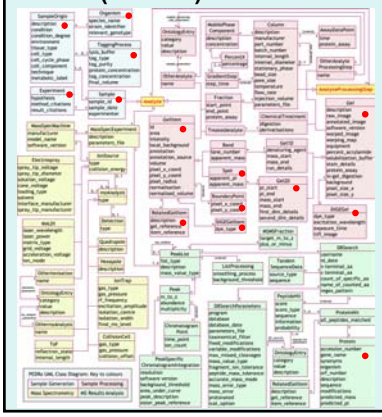
RDBMS and caches

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

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

MIAPE (PEDRo) UML Schema n-D Data Classes



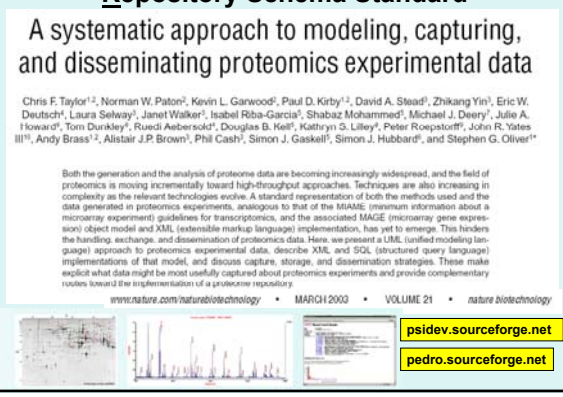
- Classes that could be used with 2D-gels

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

'PEDRo' - Proteomic Experiment Data Repository Schema Standard



A systematic approach to modeling, capturing, and disseminating proteomics experimental data

Chris F. Taylor[1,2], Norman W. Paton[2], Kevin L. Garwood[2], Paul D. Kirby[1,2], David A. Stead[3], Zhikang Yin[3], Eric W. Deutsch[4], Laura Selway[3], Janet Walker[3], Isabel Riba-Garcia[5], Shabaz Mohammed[5], Michael J. Deery[5], Julie A. Howard[6], Tom Dunkley[6], Ruedi Aebersold[4], Douglas B. Kell[6], Kathryn S. Lilley[6], Peter Roepstorff[7], John R. Yates III[10], Andy Brass[1,2], Alistair J.P. Brown[3], Phil Cash[3], Simon J. Gaskell[5], Simon J. Hubbard[8], and Stephen G. Oliver[1*]

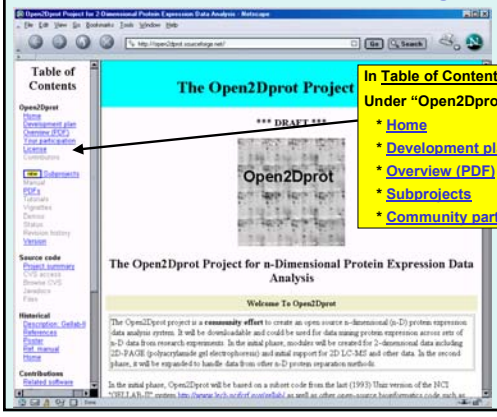
Both the generation and the analysis of proteome data are becoming increasingly widespread, and the field of proteomics is moving incrementally toward high-throughput approaches. Techniques are also increasing in complexity as the relevant technologies evolve. A standard representation of both the methods used and the data generated in proteomics experiments, analogous to that of the MIAME (minimum information about a microarray experiment) guidelines for transcriptomics, and the associated MAGE (microarray gene expression) object model and XML (extensible markup language) implementation, has yet to emerge. This hinders the handling, exchange, and dissemination of proteomics data. Here, we present a UML (unified modeling language) approach to proteomics experimental data, describe XML and SQL (structured query language) implementations of that model, and discuss capture, storage, and dissemination strategies. These make explicit what data might be most usefully captured about proteomics experiments and provide complementary routes toward the implementation of a proteomic repository.

www.nature.com/naturebiotechnology • MARCH 2003 • VOLUME 21 • nature biotechnology

psidev.sourceforge.net

pedro.sourceforge.net

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



In Table of Contents, see:

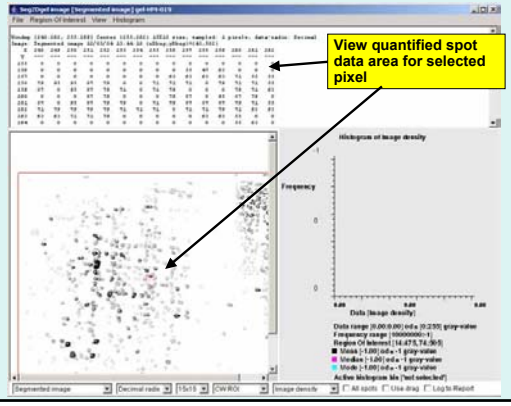
Under "Open2Dprot"

- Home
- Development plan
- Overview (PDF)
- Subprojects
- Community participation

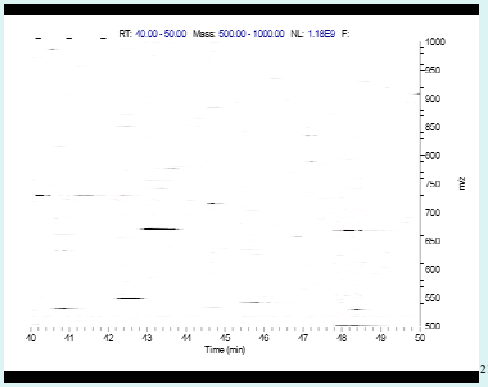
Status: Open2Dprot Pipeline Subprojects

03-02-2005

Example: Seg2Dgel Image Viewer - segmented image

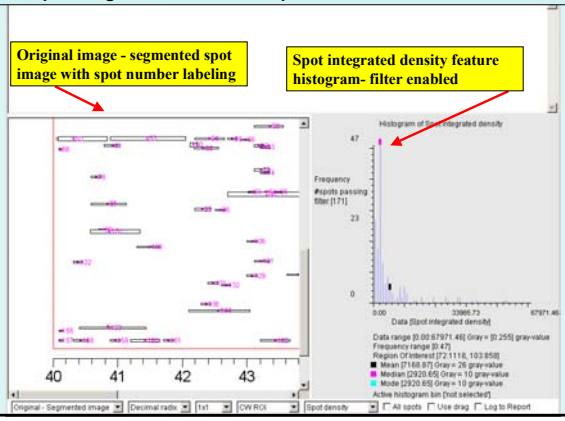


Example: 2D-LC-MS-HR low-resolution image (Veenstra/Conrads)

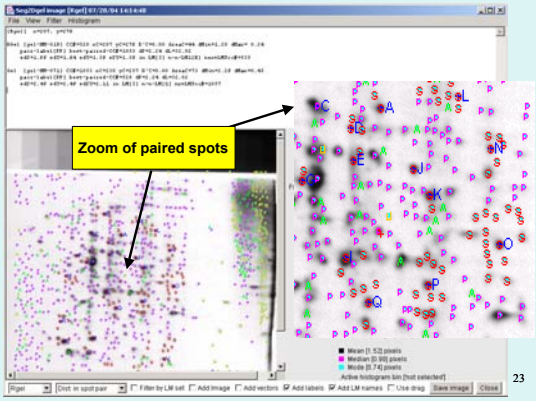


21

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



Example: CmpSpots – paired spots between Rsample and Sample



23

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 project 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

24