Open2Dgel
The Open 2D-Gel Proteomics Project

Peter F. Lemkin
Lab. Experimental and Computational Biology
National Cancer Institute
Frederick, MD, USA
lemkin@ncifcrf.gov

Presented at HUPO World Congress, Montreal, CA, Oct 8-11, 2003.
Revised: 10-23-2003

Definition of The Open2Dgel Project

The Open2Dgel project is an open-source project for the development of 2-dimensional polyacrylamide electrophoresis gel (2D-PAGE) exploratory data analysis bioinformatic tools for analyzing quantified protein expression profiles across multiple 2D gel samples from research experiments.

The tools could be adapted for use with other quantified protein separation data sources besides 2D gel data.

Introduction

- The Open2Dgel project is a community effort to create an open source 2-dimensional polyacrylamide gel electrophoresis data analysis system.
- It could be used for data mining protein expression across sets of gel samples from researcher's experiments to investigate and find significant protein expression from multiple experiments used to construct their database.
- The initial focus of Open2Dgel will be to provide an integrated set of software tools for 2D gel database analysis.
- Open source software may be freely downloaded - both executable binaries and the source code, modified and redistributed.
- Later, Open2Dgel could be expanded to handle protein expression data from other protein separation methods.

Hosting Open2Dgel on the Web

- Initially, it will be hosted and developed on the open source SourceForge.Net repository at **open2dgel.sourceforge.net**.
- It will use the same open source methodology we used in our MAExplorer **maexplorer.sourceforge.net** DNA microarray data mining software.
- A preliminary Web site, **www.lecb.ncifcrf.gov/Open2Dgel**, discusses the Open2Dgel software development plan.
- Open2Dgel could later reside as part of a general **HUPO.org** analysis sub-Web site integrated with other tools relating to protein mass spectrometry, protein arrays, Internet proteomic databases and other technologies covering a broad range of protein expression.

Overall Development Plan

- Open2Dgel will be written in Java and R languages using XML and a MySQL RDBMS - modular open source technologies aiding portability and extensibility.
- In the initial phase, **Open2Dgel** will be derived from the parts of NCI GELLAB-II system - the C-language/Unix/X-windows 1993 version (**www.lecb.ncifcrf.gov/gellab**), code from other open source proteomics and bioinformatics projects, and leverage Java/R code from MAExplorer.
- In the second phase, it could be extended with other donated 2D gel analysis and related proteomics software codes as well as developer efforts donated by the research community.

Detailed Development Plan (cont.)

- We will work with proteomics standardization groups (HUPO, PSI, MIAPE - formerly PEDRo, and others) to use a standard data database schema.
- We will encourage the user community to help expand, extend and integrate the basic paradigm with other related protein separation methods or data analysis systems using the standard proteomics schema.
- We welcome suggestions for modifying this agenda for Open2Dgel as well as bioinformatics developers offering to help with the project.

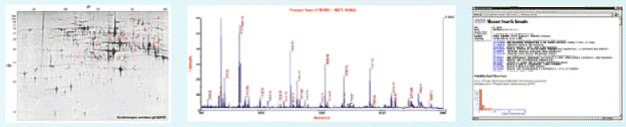
PEDRo - Proteomic Experiment Data Repository Schema

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

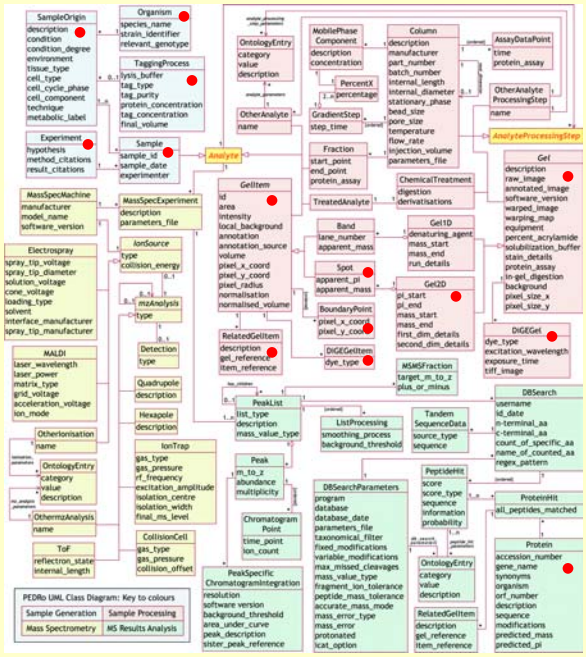
Chris F. Taylor[1,2], Norman W. Paton[2], Kevin L. Garwood[2], Paul D. Kirby[1,2], David A. Stead[3], Zhikang Yin[3], Eric W. Deutsch[4], Laura Selway[3], Janet Walker[3], Isabel Riba-Garcia[5], Shabaz Mohammed[5], Michael J. Deery[7], Julie A. Howard[8], Tom Dunkley[8], Ruedi Aebersold[4], Douglas B. Kell[5], Kathryn S. Lilley[8], Peter Roepstorff[9], John R. Yates III[10], Andy Brass[1,2], Alistair J.P. Brown[3], Phil Cash[3], Simon J. Gaskell[5], Simon J. Hubbard[6], 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 proteome repository.

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



MIAPE (PEDRo) UML Schema - 2D Gel Classes



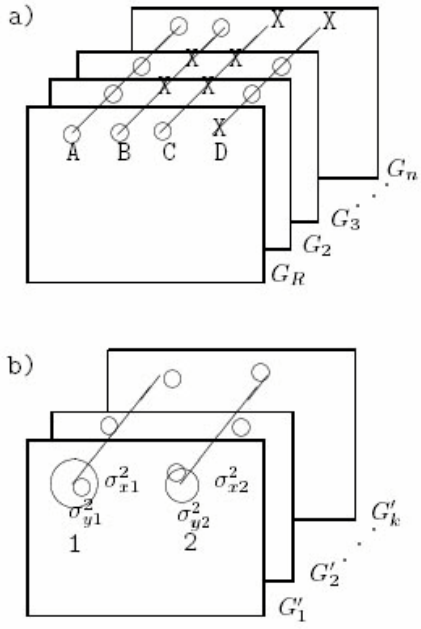
- **Classes that could be used with Open2Dgel**

Additional fields / classes are needed for Open2Dgel

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

Composite 2D Gel Database Paradigm



2D gel composite gel database (CGL) consisting of a set of n gels G_1, G_2, …,G_n with representative gel $G_r = G_1$

A canonical 2D gel database is a statistical representation of the CGL spot geometry and quantification that could be used for data mining

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

Basic Open2Dgel Analysis Pipeline

1. Accession gel sample images and experiment information — *interactive*

XML

2. Segment gel images to quantify spots for all gels — *batch*

XML

3. Create a landmark database between reference gel and remaining gels - if needed by spot pairing algorithm — *interactive*

XML

4. Pair spots between a reference gel and the rest of the gels — *batch*

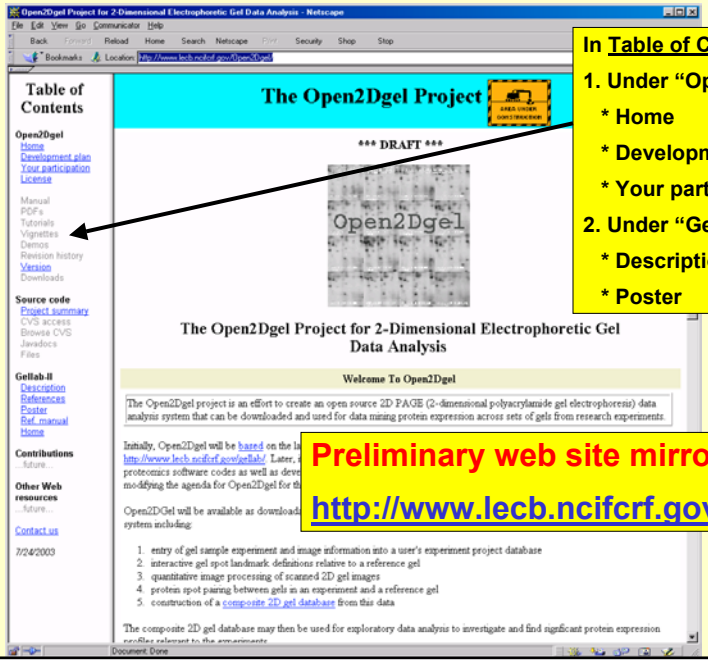
XML

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

RDBMS and cache

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

Future Home: http://open2dgel.sourceforge.net/



In Table of Contents, see:

1. Under "Open2Dgel"
   * Home
   * Development plan
   * Your participation
2. Under "Gellab-II"
   * Description of old GELLAB-II
   * Poster

Preliminary web site mirror:

http://www.lecb.ncifcrf.gov/Open2Dgel

The Initial Open2Dgel Data mining Tools

- Accession 2D gel scanned image and experiment data
- Quantify spots from gel images
- Pair spots between gels and a reference gel
- Construct composite gel database for exploratory data analysis
- Handle multiple gel samples in a database
- Manage named subsets of proteins in the database
- Manage replicate gel samples, named condition sets of samples, lists of condition sets
- Analyze data for multiple conditions expression profiles
- Data filter protein sets by statistics, clustering, set membership
- Direct-manipulation of data in graphics, spreadsheets and sample management
- Integrate R language statistical, clustering and other methods
- Integrate access to Internet proteomic/genomic data servers for user-specified protein sets

Bioinformatics Community Support Required for The Open Source Project

1. The initial effort: developers will be needed to refactor a) code from the NCI-GELLAB-II system (C/Unix/X-windows), and b) other code to the modular (Java/R/XML/MySQL-RDBMS) paradigm.

2. A few senior developers interested in taking on managerial and design roles (a long-term goal is to have multiple "project managers" in various proteomics specialties).

3. Active research groups to beta-test system with their 2D gel data

4. Help with subsequent extension/integration with other protein separation methods software/databases (mass spectrometry, protein microarrays, dye multiplexing, statistics, data mining, etc).

5. Contributions of alternative computation modules for analysis pipeline - e.g., spot quantification, pairing, statistical analysis, etc.

Summary

- The Open2Dgel project is fully open source and will be available at **http://open2dgel.sourceforge.net/** when released.

- **The project will proceed if:**

1. There is sufficient need for an open source extensible proteomics exploratory analysis tool.

2. There is sufficient interest from the research community.

3. Members of the research community are willing to work on various aspects of the project.