PM12: ISMB2006 Tutorial

Title: Automatic online data integration pipelines with Expression Profiler for bioinformatics programmers.

Topic Area:

- Database and Data Integration

Main Presenter:

- Mr Misha Kapushesky
- Microarray Informatics Team
- EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK
- ostolop@ebi.ac.uk
- +44 1223 494 647 (work), +44 7737 990 161 (cell)
- +44 1223 494 468
- http://www.ebi.ac.uk/~ostolop

- 6+ years of teaching experience: at undergraduate & graduate level
- Selected tutorials/courses in microarray informatics:
  - National Bioinformatics Network: Microarrays Module, University of Western Cape, South Africa, Sep. 20-26, 2005.
  - University of Camerino Summer School: Microarray Technology and Bioinformatics, Camerino, Italy, August 29 – Sep. 2, 2005.
  - European Networking Summer School: Data analysis approaches in microarray experiments, August 26, 2005
  - Bioinformatics Week at Purdue University: Microarray Informatics at the EBI, May 16-20, 2005.
  - EMBO Practical Course on Analysis and Informatics of Microarray Data (co-organiser): April 3-9, 2005.
  - EMBO Practical Course in Microarray Technology: Genome – Proteome – Function, May 29 - June 5, 2004, EMBO, Heidelberg, Germany. Subjects taught: “Analysis of DNA microarray data sets, data mining techniques.”
  - Joint A-IMBN/EMBO Course: Microarray Techniques: Applications in Bio-Medical Research, March 14 – 21, 2004, Tokyo, Japan. Subjects taught: “Tools for annotation of microarray experiments, missing data estimation methods, identification of differentially expressed genes (statistical tests, power analysis), clustering analysis methods, dimensionality reduction techniques.”
  - Industry Programme Workshop: Microarrays & Data Mining December 8 - 9, 2003 (Scientific Coordinator). Subjects taught: “Clustering analysis methods, identification of differentially expressed genes, Expression Profiler: Next Generation tutorial.”
  - NETTAB 2003 Workshop: Bioinformatics for the Management, Analysis and Interpretation of Microarray Data, November 27 - 28, 2003, CINECA, Bologna, Italy. Subjects taught: “Expression Profiler: Next Generation Tutorial.”
  - EMBO Practical Course on Analysis and Informatics of Microarray Data, March 16-22, 2003, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. Subjects taught: “Microarray Data Analysis Practical Session: Data Mining.”

- Industry Programme Workshop: Microarrays and Data Mining December 10 - 11, 2002, EBI, Hinxton, Cambridge UK. Subjects taught: "Using Expression Profiler for Microarray Data Analysis."
- EMBO Course on Microarrays, EMBO, Heidelberg, Germany, June 1 - 8, 2002. Subjects taught: "Using Expression Profiler for Microarray Data Analysis and Mining."

50-word abstract:
This tutorial is for bioinformaticians with programming experience who are interested in making use of EBI's resources for running complex, computationally demanding microarray data analysis pipelines via Expression Profiler. Participants will build workflows from available components (normalization, annotation, visualization, etc.) and integrate external tools (e.g., Bioconductor packages) into their analysis.

Tutorial level: Advanced

Prior knowledge required: The participants are expected to be proficient in Perl programming language, and preferably (though a brief overview will be included in the tutorial) should understand the basic techniques of microarray data analysis, such as data pre-processing, clustering, and so on. Familiarity with Bioconductor will be highly beneficial.

Suitability of this tutorial for ISMB:
As microarray datasets are entering the era of high-density array platforms, the sizes of the datasets are growing by orders of magnitude. Hardware requirements in terms of computational power and memory are growing as well. This tutorial will allow advanced bioinformaticians who are working in the area of functional genomics data analysis to access the powerful infrastructure provided by the EBI (via the Expression Profiler platform) by composing data analysis pipelines. Expression Profiler provides a number of standard best-practice components for data pre-processing, annotation and visualization but, more importantly, the participants will learn how to incorporate other cutting-edge algorithms (e.g. those in the open source Bioconductor toolbox) into their analysis.

This tutorial directly addresses the need for rapid development and prototyping of powerful, flexible data integration strategies. The proposed workshop targets the advanced bioinformaticians who have developed their own data analysis routines, perhaps contributed to Bioconductor and are looking for an integrated, uniform way to place their analyses within larger data analysis pipeline, taking advantage of developed data management and processing technologies.

Profile of Presenter
On completion of a degree in mathematics from Cornell University, and after a year as visiting graduate student at Oxford University where he developed computer-simulation models of extracellular signalling pathways, Mr Kapushesky worked in the software industry, including several high-tech start-ups, developing novel data integration interfaces, large-scale databases and data analysis algorithms.

Joining the European Bioinformatics Institute in the Microarray Informatics Team, his main project has been the development of Expression Profiler, a major on-line

platform for exploratory data analysis and visualization, which comprises numerous components for performing all the basic steps involved in building large-scale functional genomics data comprehension pipelines. In addition to building interfaces to a number of well-known algorithms, he has participated in novel algorithm and data visualization methods development. He is currently leading the software engineering effort of the Microarray Informatics Team in terms of data integration and interface development across all our services. Mr Kapushesky is a co-organiser and a regular instructor of regular EMBO Microarray Informatics practical courses and a visiting lecturer in numerous data analysis and integration workshops.

Current projects include participation in the EU DIAMONDS project, where the Expression Profiler framework is at the centre of developing a portal for storing, analysing and modelling high-throughput cell-cycle genomic data, participation in the BioMap data warehousing and integration project, coordinated at University College London, and a collaboration with Imperial College on a grid-supported microarray data analysis and storage management system.

Tutorial Outline:

The overall goals of this tutorial are trifold:

1) Introduce users to basic exploratory data analysis possible through Expression Profiler and its integrated access to several databases: the ArrayExpress Repository, the ArrayExpress Data Warehouse, BioMart and others.
2) Teach participants to automate basic data analysis pipelines, execute them automatically on EBI's servers and construct more advanced data integration workflows.
3) Allow participants to integrate additional algorithms into the presented framework in order to execute and publish customised data analysis workflows.

Part I: Introduction to Expression Profiler (40 mins)

- This section will provide a brief tutorial on using Expression Profiler for basic exploratory data analysis and visualisations. The participants will interactively apply several basic microarray data analysis components available within the EP framework.
    - Data loading & pre-processing, automatic integration with ArrayExpress - public microarray data repository
    - Automatic data re-annotation via integration with BioMart
    - Data normalisation algorithms and quality control components
    - Exploratory analysis & visualization: clustering, dimensionality reduction methods

Part II: Automating Expression Profiler (80 mins)

- In this section the participants will learn how to automate and script Expression Profiler to execute the simple data analysis workflow from the previous part without any human interaction.
    - Introduction to Web Services
    - Programmatic execution of individual EP components
    - Multiple dataset management with Expression Profiler
    - Chaining components via Web Service XML queries

- Multiple database access and data retrieval
- Visual workflow design

Part III: Integrating Novel Algorithms (90 mins)

- In this section the participants will build a more complex data analysis pipeline that will incorporate novel algorithms not available in the basic Expression Profiler set of components.
  - Rapid application development: building a custom Expression Profiler component
  - Integrating Bioconductor packages: data analysis and visualization
  - Interfacing with external bioinformatics Web Services (Taverna, BioMoby, etc.)
  - Saving, publishing and accessing data analysis results

Part IV: Discussion (30 mins)

- The participants will have a chance to ask questions or try building their own custom workflows.