Accessible Reproducible Research

Michael Reich
Director of Cancer Informatics Development
Broad Institute of MIT and Harvard

February 19, 2011
Overview

• Background: Molecular Classification of Cancer

• Challenges to Reproducible Research

• GenePattern: A platform for reproducible genomic research

• GenePattern Reproducible Research Document
Background:
Molecular Classification of Cancer

Golub et al. *Science* 1999
Cancer Classification Workflow

Dataset

8 analysis tools/scripts
10 analysis steps

A reproducible research environment is required to support multi-tool, multi-dataset research
Challenges of Reproducible Research

• Claerbout (2000) – “An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”

• How did you get those results?
  • New versions of methods, algorithms, implementations
  • Varying parameter settings
  • Same data called different things by different “owners”
  • Same name for different objects in different repositories or organisms

A Reproducible Research Environment

• Every new paper should provide sufficient information to completely reproduce the results:
  • Data
  • Tools or programs
  • The means to execute the programs on the data without specialized knowledge
Reproducible Research Environment: GenePattern

Module Repository
- KNN
- GISTIC
- SVM
- FLAME
- PCA
- GSEA
- NMF
- CBS

~150 modules

Module Integrator

Pipeline Environment
- all_aml_train
  - Preprocess
    - SOM Clustering
    - Class Neighbors
  - SOM Cluster Viewer
- all_aml_test
  - Preprocess
    - Weighted Voting Cross-Val
    - Weighted Voting Train/Test
  - Prediction Results Viewer

Golub and Slonim et. al 1999

Client User Interfaces
- Visualizer
- Web
- Programming

KNN
GISTIC
SVM
FLAME
GSEA
PCA
NMF
CBS

~150 modules

Preprocess
Class Neighbors
Weighted Voting
Train/Test
Gene Pattern Interface
GenePattern Community

GenePattern Users
- First released in 2004
- Current release: version 3.2.3 6/2010
- Over 15,000 registered users
- Users mailing list: over 2000 members
- www.genepattern.org

Project collaborations
- TGCA: The Cancer Genome Atlas Project
- ICBP: Integrative Cancer Biology Program
- NCBC: National Centers for Biomedical Computing
- caBIG: cancer Biomedical Informatics Grid
- Stanford University: SMD
- European Nutrigenomics Organization (NuGO)
Provenance in GenePattern

• All analyses are recorded
  • Input files
  • Parameters
  • Code version (via LSID)

• Any previous analysis can be reloaded and rerun
• Pipelines can be cloned, edited, imported/exported
• Each pipeline edit creates a new version
• Pipelines can be inferred from result files, eliminating the need to know all the steps beforehand
Previous Approaches to Reproducible Research Documents

• Don Knuth: Literate Programming
• Claerbout: use of Makefiles
• Buckheit and Donoho: plots should be reproducible
• Duncan Temple Lang: extensible dynamic documents
• Tony Rossini: Literate Data Analysis
• Fritz Leisch: Sweave
• Robert Gentleman: Compendium combining R & LaTeX

Adapted from Robert Gentleman
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

Todd R. Golub\textsuperscript{1,6*}, Donna Slonim\textsuperscript{1,7}, Pablo Tamayo\textsuperscript{1}, Christine Huard\textsuperscript{1}, Michelle Gassenbeek\textsuperscript{1}, Jill Mesirov\textsuperscript{1}, Hilary Coller\textsuperscript{1}, Mignonne Loh\textsuperscript{1}, James R. Downing\textsuperscript{1}, Clara D. Bloomfield\textsuperscript{1}, and Eric S. Lander\textsuperscript{1, 2, 6}

\textsuperscript{1}Whitehead Institute/MIT Center for Genome Research, Cambridge, MA 02142
\textsuperscript{2}Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115
\textsuperscript{3}St. Jude Children’s Research Hospital, Memphis, TN 38105
\textsuperscript{4}Comprehensive Cancer Center and Cancer and Leukemia Group B, The Ohio State University, Columbus, OH 43210
\textsuperscript{5}Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139
\textsuperscript{6}\textsuperscript{*Corresponding author. Email: t.golub@broadinstitute.org}
Molecular Classification of Cancer:  
Class Discovery and Class Prediction  
by Gene Expression Monitoring

Todd R. Golub\textsuperscript{1,2,4,*}, Donna Slonim\textsuperscript{1,*}, Pablo Tamayo\textsuperscript{1}, Christine Huard\textsuperscript{1},  
Michelle Gasenbeck\textsuperscript{3}, Jill Mesirov\textsuperscript{1}, Hilary Collet\textsuperscript{1}, Mignon Loh\textsuperscript{3},  
James R. Downing\textsuperscript{2}, Clara D. Bloomfield\textsuperscript{4}, and Eric S. Lander\textsuperscript{1, 2, 4}

\textsuperscript{1}Whitehead Institute/ MIT Center for Genome Research, Cambridge, MA 02142  
\textsuperscript{2}Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115  
\textsuperscript{3}St Jude Children’s Research Hospital, Memphis, TN 38105  
\textsuperscript{4}Comprehensive Cancer Center and Cancer and Leukemia Group B, The Ohio  
State University, Columbus, OH 43210
protein important in cell adhesion in fibroblasts, but a role in hematopoiesis has not been previously reported.\(^{(2)}\)

![Heatmap showing expression levels of the 50 informative genes from 58 acute leukemic samples, most highly correlated with ALL/AML class distinction. Each row corresponds to a gene, and the columns represent samples, colored from low to high normalized expression.](image-url)
Appendix

Retrun of Fig. 3. B. added at 3/24/2008 4:51:48 PM

User: John Smith. Retrun executed with default parameters and datasets.
Molecular Classification of Cancer: 
Class Discovery and Class Prediction 
by Gene Expression Monitoring

Todd R. Golub\textsuperscript{1,2,5,6}, Donna Slonim\textsuperscript{1,5}, Pablo Tamayo\textsuperscript{1}, Christine Huard\textsuperscript{1}, 
Michelle Gassenbeck\textsuperscript{1}, Jill Mesirov\textsuperscript{1}, Hilary Coller\textsuperscript{1}, Mignon Loh\textsuperscript{2}, 
James R. Downing\textsuperscript{2}, Clara D. Bloomfield\textsuperscript{4}, and Eric S. Lander\textsuperscript{1, 5, 6}

\textsuperscript{1}Whitehead Institute/ MIT Center for Genome Research, Cambridge, MA 02142 
\textsuperscript{2}Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115 
\textsuperscript{3}St. Jude Children’s Research Hospital, Memphis, TN 38105 
\textsuperscript{4}Comprehensive Cancer Center and Cancer and Leukemia Group B, The Ohio State University, Columbus, OH 43210 
\textsuperscript{5,6}
Potential Future Directions

- PDF implementation
- Cross-platform MS Word implementation
- Interoperability with Web-based publication resources, e.g. myExperiment
- Inference engine-based assertion of dataset constraints
Acknowledgements

• Helga Thorvaldsdottir
• Barbara Hill
• Jon Bistline
• Ted Liefeld
• Peter Carr
• Marc-Danie Nazaire

• Microsoft
• Infusion
• Persistent Systems

PI Jill P. Mesirov

http://genepatternwordaddin.codeplex.com
http://www.genepattern.org
http://www.genomespace.org