GGobi : Interactive and dynamic data visualization system

Bioinformatics and Biostatistics Lab.,
Seoul National Univ. Seoul, Korea
Eun-Kyung Lee

Outline

- interactive and dynamic graphics
- Exploratory data analysis and Data mining
- What is GGobi?
- Main features of GGobi
- Demo with a couple of examples

Interactive vs. Dynamic Graphics

- Interactive graphics
  - a user can actively manipulate the visual graphics by input devices and make changes based on the visual result.

- Dynamic graphics
  - the visible graphics change on the computer screen without further user interaction

--> Interactive and dynamic graphics

Interactive and dynamic graphical methods

- Focusing
  - zooming, slicing, rescaling, reformatting
- Arranging
  - rotation, grand tour, guided tour, manual tour
- Linking
  - Linked brushing and identification

EDA vs. DM

- Exploratory Data Analysis (EDA)
  - numerical or graphical detective work
  - a continuation of Tukey's idea to use graphics to find structure, general concepts, unexpected behavior, etc. in data sets by looking at the data.

EDA vs. DM

- Data Mining (DM)
  - Data mining is exploratory data analysis with little or no human interaction using computationally feasible techniques - Wegman
- Visual Data Mining (VDM)

= DM + Statistical graphics

What is GGobi?

- A direct descendent of XGobi
- A data visualization system with interactive and dynamic methods for the manipulation of views of data.
- provide various plots with multiple plotting windows system
- use XML file format for data
- can be easily extended, either by being embedded in other SW or by the addition of plugins
- able to use in R (rggobi)

GGobi's main features

1. Appearance
   - Use GTK+
   - single session can support multiple plots
   - single process can support multiple independent session
   - support several types of plots
     - scatter plot, parallel coordinate plot, scatter plot matrix, time series plot, barchart
   - include interactive tools to specify and tune color maps
   - able to add variables on the fly
   - panning and zooming

GGobi's Main features

2. Portability
   - runs under various platforms, like Linux, Windows or Mac.
3. Data format
   - XGobi : use several files (.dat, .col, .row, .glyphs, .colors, etc)
   - use XML
   - allow complex characteristics and relationships in data to be specified
   - multiple dataset can be entered in a single XML file and specifications can be included for linking them

GGobi's Main features

4. Embedding in other SW
   - GGobi can be treated as a C library and directly embedded in other SW, then controlled using an application program interface (API)
   - This allows GGobi functionality to be integrated into one's own stand-alone application and provide as an add-on to existing language and scripting environments
5. Extending with plugins
   - The plugin mechanism allows to provide add-on extensions to GGobi that are not part of the core design
   - data viewer, ggvis (MDS), Variogram Cloud, Save Display Description

GGobi : File

- open
  - XML from files
  - XML from URL
  - CSV
- New
  - open new session
- Save
  - as XML : keep all the information including color, glyph, etc.
  - as CSV : keep only numeric data values

GGobi : Display

- open new plot window

- New Scatterplot Display
- New Scatterplot Matrix
- New Parallel coordinates Display
- New Time Series
- New Barchart

GGobi : View

- 1D plot
- XY plot : 2D plot
- 1D tour : project data into 1D space
- Rotation : use three variables
- 2D tour : project data into 2D space
- 2x1D tour : use 2 different 1D tour

GGobi : interaction

- Scale
- Brush
- Identify
- scale
- EditEdges : add edges or add points in Display
- MovePoints : move points in Display

GGobi : Tools

- Variable Manipulation ; Variable Transformation
- Sphering (PCA)
- Variable jittering : prevent point mass viewing
- Color Schemes; Automatic Brushing
- Color & Glyph groups; Case Subsetting & Sampling
- Missing Values
- plugins
  - Data Viewer, ggvis(MDS), Variogram Cloud, Save Display Description

RGGobi

- able to use GGobi in R
- Link, including programming customized GUIs containing linked GGobi plots, writing new linking rules in S, and responding to GGobi events, create GGobi plugins written in R.

Example 1 : Restaurant Tipping

- In early 1990, one waiter recorded information about each tip he received over a period of a few months working in one restaurant. He collected several variables; (n=244)

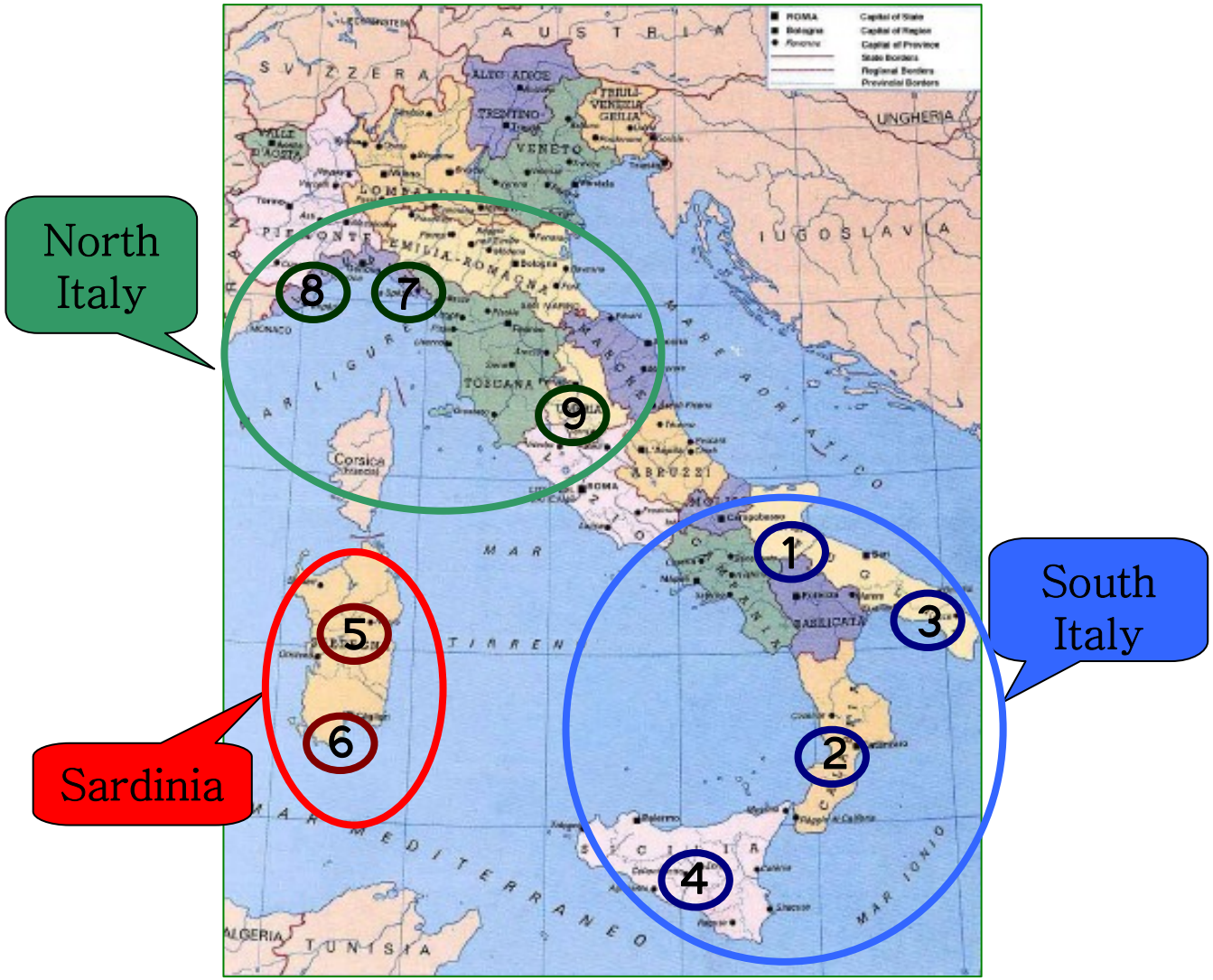
  - TOTBILL : total bill in dollars
  - TIP : tip in dollars
  - SEX : gender of the bill payer; male(0), female(1)
  - SMOKER : whether the party included smokers or not
    No(0), Yes(1)
  - DAY : days of week ; Thu(3), Fri(4), Sat(5), Sun(6)
  - TIME : lunch(0), dinner(1)
  - SIZE : size of the party

Example 2 : Italian Olive Oils

- This data consists of the percentage composition of 8 fatty acids found in the lipid fraction of 572 Italian olive oils (n=572)

  - Region : South(1),North(2) or Sardinia(3)
  - Area : North-Apulia(1), Calabria(2), South Apulia(3), Sicily(4), Inland Sardinia(5), Costal Sardinia(6), East Liguria(7), West Liguria(8), and Umbria(9)
  - Palmitic
  - Palmitoleic
  - Stearic
  - Oleic
  - Linoleic
  - Linolenic
  - Arachidic
  - Eicosenoic

Italy


North Italy
South Italy
Sardinia
1
2
3
4
5
6
7
8
9
AUSTRIA
SVIZZERA
UNGHERIA
JUGOSLAVIA
VENETO
TOSCANA
ABRUZZI
BASILICATA
Corsica
MAR TIRRENO
MAR MEDITERRANEO
MAR IONIO
ALGERIA
TUNISIA

Example 3 : Leukemia data

n = 72 : # of observation

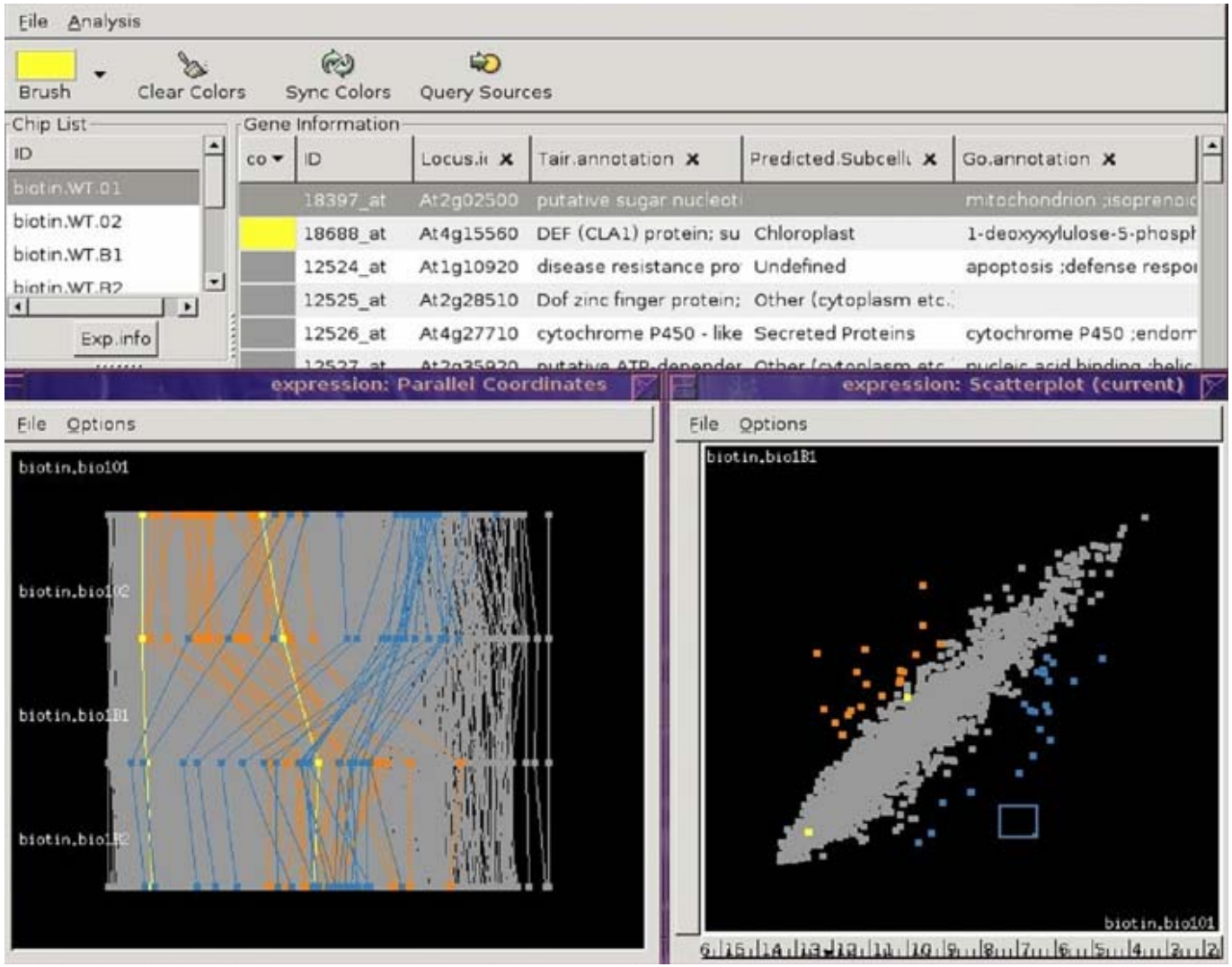
p = 3571 → p=40 : # of genes

- acute myeloid leukemia(AML) : 25 cases
- acute lymphoblastic leukemia(ALL) : 47 cases
  - ◉ B-cell ALL : 38 cases
  - ◉ T-cell ALL : 9 cases

exploRase

- Visual data mining tools for microarray data and metabolic networks
- Visual data analysis interface for microarray data and metabolic networks
- Based on R and GGobi
- Provide EDA tool using direct manipulation
- analyze the connections between microarray data and metabolic pathway visually and interactively
- combine statistical analysis results with interactive plots to improve the analysis.

exploRase


File Analysis
Brush
Clear Colors
Sync Colors
Query Sources
Chip List
ID
biotin.WT.01
biotin.WT.02
biotin.WT.B1
Exp.info
Gene Information
ID
Locus.ic
Tair.annotation
Predicted.Subcell
Go.annotation
18397_at At2g02500 putative sugar nucleoti mitochondrion ;isoprenoic
18688_at At4g15560 DEF (CLA1) protein; su Chloroplast 1-deoxyxylulose-5-phospl
12524_at At1g10920 disease resistance pro Undefined apoptosis ;defense respo
12525_at At2g28510 Dof zinc finger protein; Other (cytoplasm etc.
12526_at At4g27710 cytochrome P450 - like Secreted Proteins cytochrome P450 ;endom
expression: Parallel Coordinates
File Options
biotin.bio101
biotin.bio1B1
expression: Scatterplot (current)
File Options
biotin.bio1B1
biotin.bio101

Discussion

- Full marriage between GGobi's direct manipulation graphical environment and R's familiar extensible environment for statistical data analysis

➔ powerful tool for visual data mining

* all references and documents are on the web

http://www.ggobi.org