An Insight-Based Methodology for Evaluating Bioinformatics Visualizations

Purvi Saraiya, Chris North and Karen Duca. IEEE Transactions on Visualizations and Computer Graphics. v.11 no.4 July/Aug 2005.

Meredith Pulley
INLS 706
October 16, 2006

Why are visualization tools important?

- Type of data working with--large, complex data sets to analyze (especially in biology domain)
  - microarray experiments—measures expression of hundreds or thousands of genes at once. The challenge currently facing scientists is to find a way to organize and catalog this vast amount of information into a usable form
- Ideal role visualization tools play in data analysis
  - Provide different visualizations of data
  - Provide ability to manipulate content (data)/visualizations
  - Provide method of sharing data with other researchers
    - Together, these capabilities aid in sense-making and learning process
      - Pattern recognition
      - Drawing conclusions
      - Make hypotheses to explain results, predictions/future experiments
      - Best tools: Allow for rapid interactions with data, conceptualization of results in larger context, larger implications of data in particular domain (links to public gene databases, literature databases, etc)
      - Ex. How multiple gene products work together; gene in pathway

Expectations for article

- Learn about the users:
  - How do scientists use these tools? Type of tasks want to accomplish?
  - How do scientists choose from the available tools?
    - Does type of data influence choice? How long will they spend learning a tool? Level of expertise needed to work with tool?
- Learn about the tools:
  - What features offered in visualization tools?
    - Design, visualizations offered, types of interactions available
- User + tool (User interaction with tool)
  - How do users evaluate tools:
    - Which features are perceived by users as the most useful?
    - Role of usability
    - Types of insight gained (observations, hypotheses, depth of insight—what do users actually learned)
  - What are the shortcomings of existing tools?

Study methodology

- Typical visualization studies: controlled experiments
  - Limitations
- This study: introduced method to model/capture open-ended nature of visual data exploration—”think-aloud analysis”
  - Combination of controlled experiment and usability testing methodology
  - Expected benefits of methodology

Development of methodology

- Use pilot study--key developments:
  - User-derived definition of insight (generated list of 8 characteristics of insight
  - Insight as a “unit of discovery”
    - Measurable (quantifiable)--used above list in real experiment to code these insight occurrences during participants “think-aloud” visual data analysis while using tool
    - Reproducible methodology

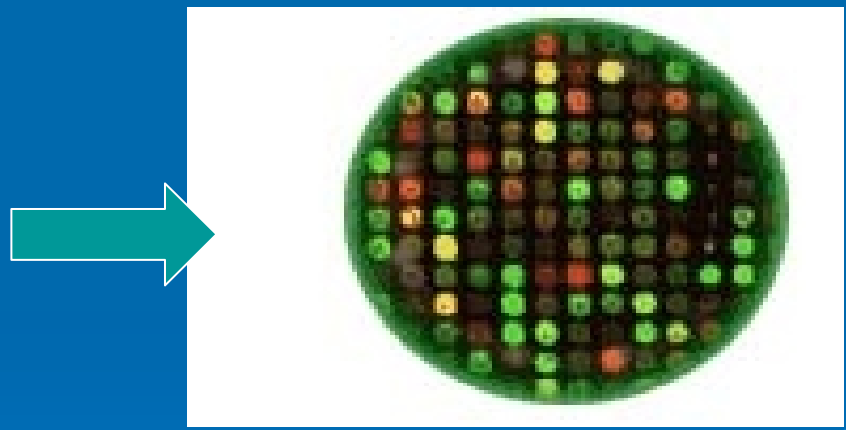
Experimental design: measuring insight gained from tools

- Objective→ Evaluation of bioinformatic visualization tools in terms of insight provided.
- Measure by individual insight occurrences and overall amount of learning
  - Quantifiable in terms of:
    - Amount of insight gained
    - Time to gain insight(s)
    - Quality (value) of insight gained (domain value)
    - Depth of finding

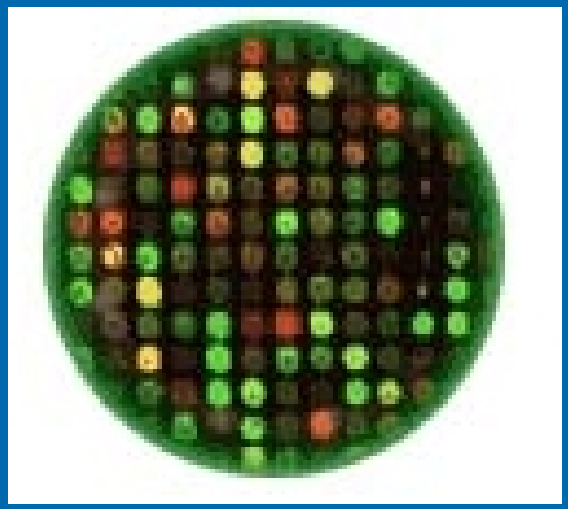
Experimental design

- Independent variables:
  - Microarray visualization tools (5) (See Table 4 and next slide):
    - Clusterview
    - TimeSearcher — Free
    - HCE
    - Spotfire
    - GeneSpring — Commercial
  - Data sets (3):
    - Timeseries data set (time points)
    - Virus data set (categorical-cells infected with one of three viral strains (measured expression of one of these variables))
    - Lupus data set (multicategorical-measured expression in control (healthy) and SLE samples)

Microarray chips



Colors of a microarray



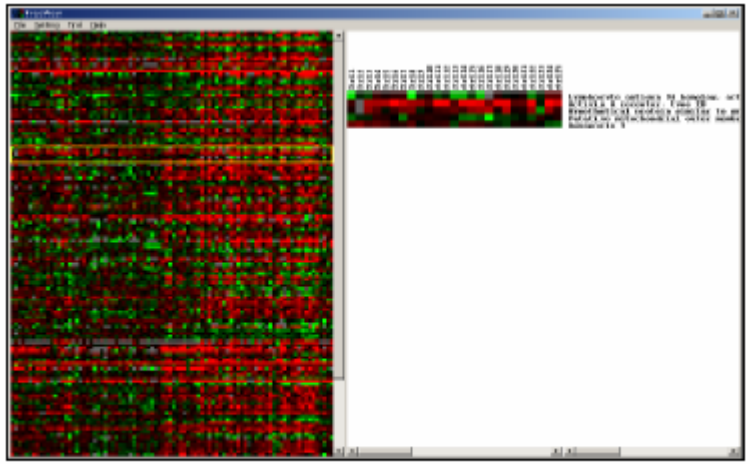


Figure 1: Cluster/Treeview (Clusterview) [9]

Each spot on an array is associated with a particular gene. Each color in an array represents either healthy (control) or diseased (sample) tissue. Depending on the type of array used, the location and intensity of a color will tell us whether the gene, or mutation, is present in either the control and/or sample DNA. It will also provide an estimate of the expression level of the gene(s) in the sample and control DNA.

Open access tools

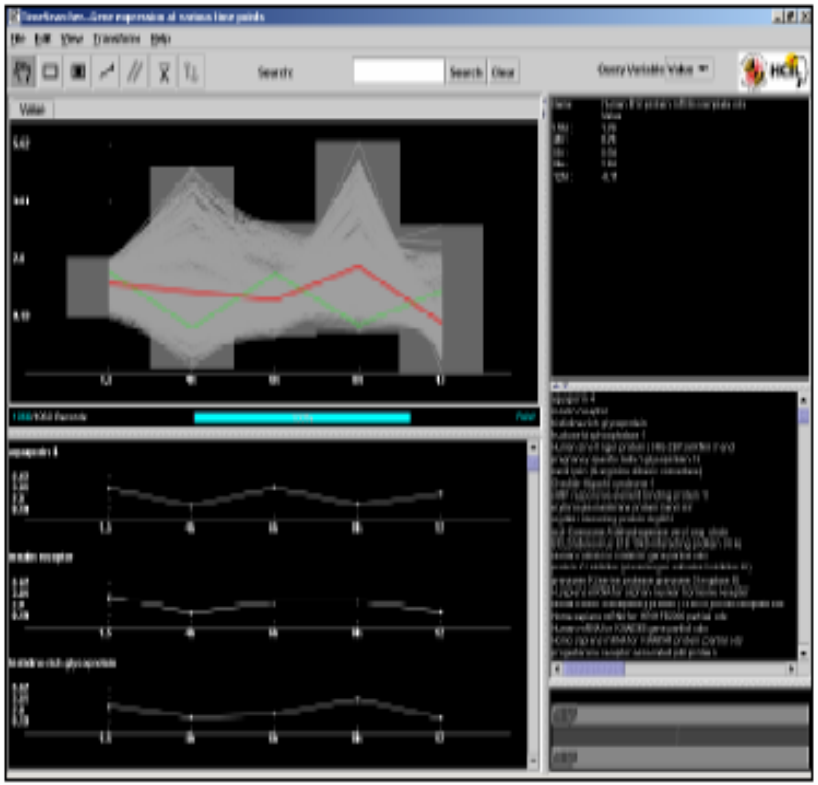


Figure 2: TimeSearcher [16]

Time series display of all data attributes

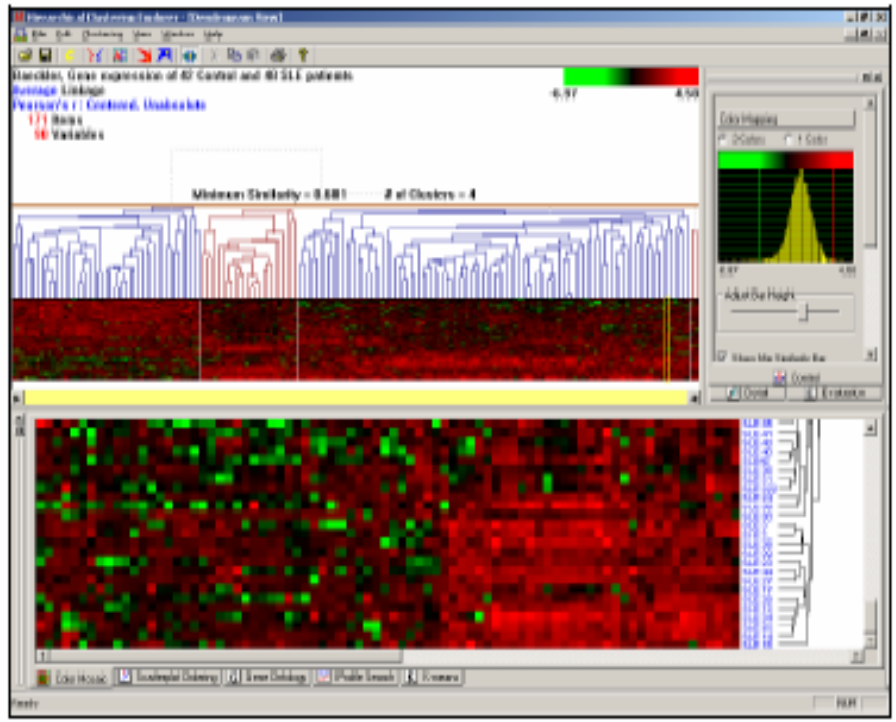


Figure 3: Hierarchical Clustering Explorer (HCE) [27]

Cluster dendogram

Commercial tools

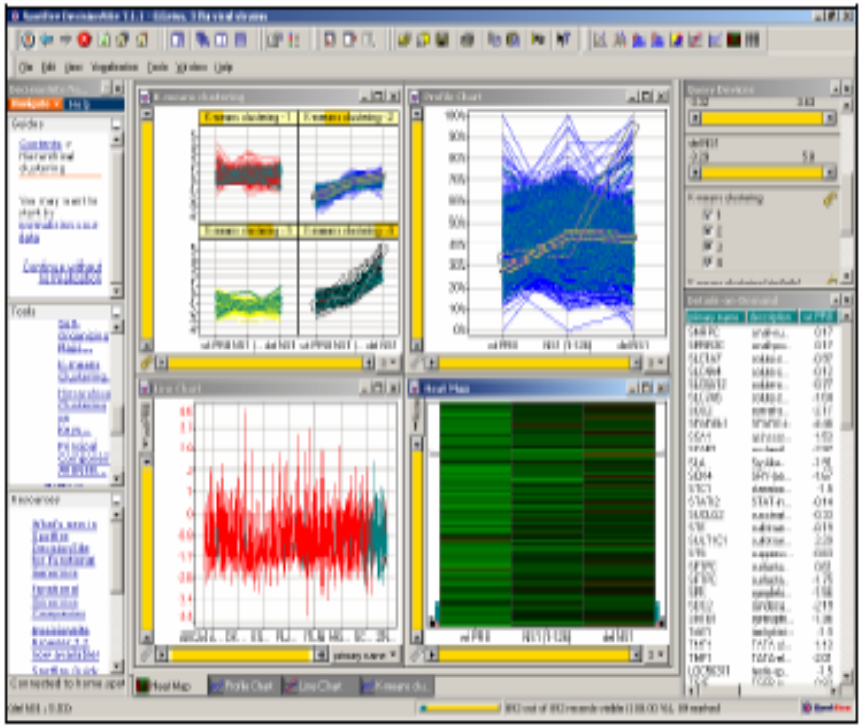


Figure 4: Spotfire® [30]

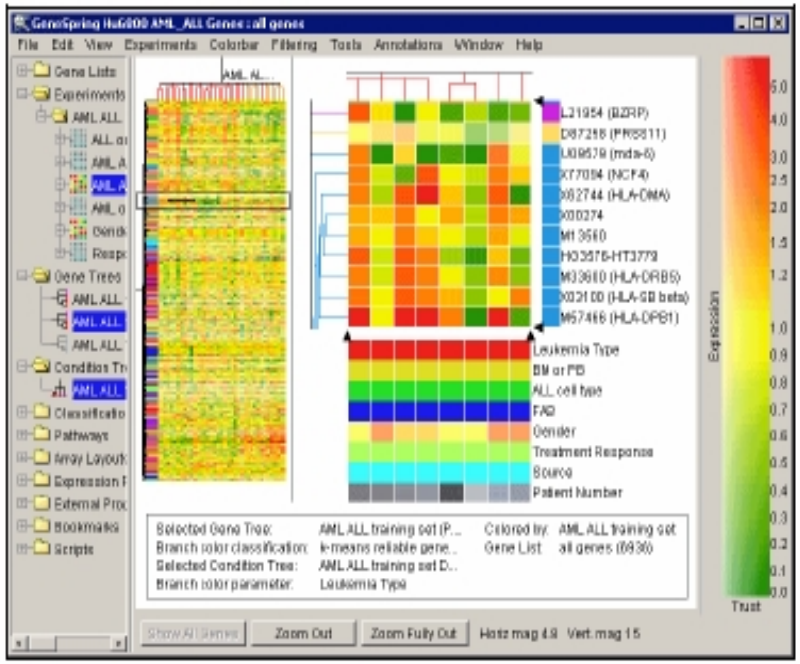


Figure 5: GeneSpring® [12]

Clustered parallel coordinates

Design: Assignment of tools

- Study population N=30; grouped by education level, professional title, experience with microarray data analysis
  - Domain Expert N=10
  - Domain Novice N=11
  - Software Developer N=9
- Controlled for user experience with tool
- 6 users per tool; 1 data set and 1 tool per user
- Procedure for participant data analysis

Presentation of Results

- Examined user insight with tool in 5 ways:
  1. Evaluation of measured insight
     - Higher value + count=more effective tool for providing insight
     - Lower time to first insight= faster learning curve for tool
     - Ideal: fastest amount of information over shortest possible time
       - Spotfire best general performance—higher insight levels at rapid insight pace.
       - Clusterview and TimeSearcher-rapid insight, then reaches limit
       - Genespring—good for overall patterns but too complicated to use
  2. Comparison of insight with tools within each data set (for data set = timeseries, viral or Lupus data set)
     - TimeSeries: Best--Spotfire and TimeSearcher
     - Viral: Best--HCE
     - Lupus data set: Best--Spotfire and Clusterview

Presentation of Results

3. Comparison of insight with tools across the 3 data sets

-Timesearcher—Best for time series

-HCE-best for Viral data set; bad for Lupus data set

-Other tools well rounded

4. Insight curves—actual vs perceived user insight over time

-Spotfire and GeneSpring users felt they gained more insight

5. User evaluations of tools

- functions users found valuable
- Visual representations and interactions
- Summary of user comments on tool

Discussion of results

Tool features and learning curves

1. Association between user insight confidence and comprehensiveness of tool (Spotfire vs. Clusterview on Lupus data set).
2. Free tools (TimeSearcher and HCE)→ Focused on specific tasks → simpler user interface→ Faster for user to learn, generate insights quickly

   --Performance is data type dependent
3. Spotfire-best overall performance

Key: Large feature set, short learning time

Discussion of results

Shortcomings of tools (user-tool interactions):

4. Tools do not adequately link data to biological meaning: Domain expertise/background had no effect on actual insight gained--performance the same among all three categories of participants (domain experts, domain novice, software developers) .

   --Only difference was in users perceived insights gained.

Results: Tool Shortcomings

So, need for tools to provide more information-rich environment—allow user (here, domain expert) to recognize patterns in data set to gain meaning in larger biological context via link to public gene databases, literature databases.

5. Usability issues: Usability of interactions outweighs user choice of visualization, even if not initial preference. Too, usability issues influences outcome performance.

6. Interaction design: Better tool support for user control over content manipulation and better integration of techniques into overall interaction model. ex. ability to select and group (cluster) genes was most common interaction users performed.

Other

7. User motivation: Low motivation for detailed analysis of visualizations= most comments were of the type “breadth” rather than depth

Learning from this study

- For biologists:
  - Visualization tools influence interpretation of data and insight gained
  - Data set dictates which tool is best to use
    - Time series=Timesearcher, Viral=HCE
  - Larger tools (Spotfire and GeneSpring) consistent across different data sets, good researchers working with multiple kinds of data
  - Spotfire best overall performance
- For visualization designers:
  - Importance of usability of interaction techniques in tool (ability to select and cluster data)
- For evaluators:
  - Importance of developing methodology to model real life situations while offering qualitative insights/explanations to quantitative results (use of insight definition and think-aloud procedure)

Limitations of study

- Short tool usage/measurement of insight period—not realistic measure--call for longitudinal study to measure long term insight
- Lack of user motivation—no ties to data, no incentive for in-depth analysis as would with users own research generated data
- Unfamiliar with data set--difficult to appreciate biological relevance of data
- Each participant was unfamiliar with tool used—could influence insight/tool use

Strengths of study

- Introduction of new methodology—user centric, more realistic than typical visualization experiments
- Varied and detailed examination of results
- Provided difficulties of approach; provision of information supports EBP
- Recognized study limitations and proposed solutions
- Generated suggestions for more meaningful tools, illuminated information needs of biologists
- Suggested Best Tool design: Offer a variety of visualization and interaction choices. Need to offer in-depth data exploration while maintaining easy usability of tool (ex. Clusterview-users thought too basic, GeneSpring-too difficult, Spotfire-impressive, comprehensive tool, but need for better usability on some visualizations)