SAS® Microarray Solution for the Analysis of Microarray Data

Susanne Schwenke, Schering AG
Dr. Richardus Vonk, Schering AG
Overview

- Challenges in Microarray Data Analysis
- Software for Microarray Data Analysis
- SAS® Scientific Discovery Solution
- SAS® Microarray Solution
- SAS® Genetic Marker Solution
- Conclusions
- SDS at Schering AG
Challenges in Microarray Data Analysis

- Pharmacogenomics is increasingly integrated into the drug development process.
- As of yet, reporting of results is voluntary except for situations where claims are derived from the findings (FDA, 2005).
- Consequently, the need for validated software for microarray data analysis is growing.
- Due to the high complexity of microarray studies, tools need a high flexibility to incorporate standard statistical approaches as well as state-of-the-art methods developed in the field of microarray analysis.
Challenges in Microarray Data Analysis

- Many problems in microarray data are well known in statistics or bioinformatics...
  - Group structure.
    - Simple parametric or non-parametric tests.
  - Clustering.
    - Hierarchical clustering.
  - Classification.
    - Logistic regression, discriminant analysis.
  - Repeated measurements.
    - Mixed models modeling time structure.
  - Time-to-event data.
    - Survival analysis.
Challenges in Microarray Data Analysis

... and others pose new challenges:

- High dimensionality.
  - The Affymetrix® HU133 Plus 2.0 microarray represents >38,500 gene sequences, the number of variables exceeds by far the number of samples.

- Correlation structures.
  - Genes are represented by multiple probe sets.
  - Pathways of genes show complex correlation structures.
  - Usually, no clear hypotheses about structures are available.

- Lack of gold standards in terms of data.
  - No real-world data sets are available, for which the 'truth' would be known.
The field of microarray data analysis is rapidly evolving and changing.

Keeping track of latest developments and assessing their quality is one of the major challenges, especially in software development.

Many available tools origin in universities and research groups. They usually solve very specific problems (e.g., Statistical Analysis of Microarrays (SAM), Tusher et al(2001)).
Software for Microarray Data Analysis

- Probably the largest collection of analysis tools is collected in the Bioconductor project (Gentleman et al (2004)), which is closely related to the R project.
  - The software is freely available and open source.
  - Everybody is free to contribute their methods including documentation to the project.
  - The Bioconductor-community is very active, therefore Bioconductor is evolving fast and is very much up-to-date.
  - The package undergoes no thorough validation, inclusion of methods is not curated.
Software for Microarray Data Analysis

- Some software developers also offer packages for microarray data analysis:
  - **SAS®** offers the Scientific Discovery Solution (SDS).
  - **Insightful®** uses S-PLUS®'s vicinity to Bioconductor and offers some Bioconductor procedures incorporated in the S+ArrayAnalyzer™.
  - **GeneData® AG** offers the Expressionist™ for microarray data analysis. It is easy to use (no programming skills needed) and emphasizes visualization. The scope of statistical methods is comparably small, but R functions can be incorporated.
  - Probably there are many more (to come)...
The SDS consists of three main blocks:

- Microarray Solution for microarray analyses.
- Genetic Marker Solution for analysis of genetic data.
- [Proteomics Solution]
SAS® Scientific Discovery Solution (SDS)

- Framework to manage data and results
SAS® Scientific Discovery Solution (SDS)

- Different analysis methods are available through the Analytical Process Selector.
- The items reflect the main steps in microarray analysis.
SAS® Scientific Discovery Solution (SDS)

- Catalog of Requirements:
  - General software, interfaces.
  - Pre-processing and transformation algorithms.
  - Univariate test procedures, implementation of multiple testing adjustments.
  - Algorithms for clustering and classification.
  - Annotation of probe sets.
  - Visualization.
  [   Methods for statistical genetics.]
Non-Clinical Statistics

SAS® Microarray Solution

- General Software, Interfaces:
  - SAS® programming language well established. +
  - Concise and detailed documentation of methods used. +
  - Generation of detailed log-files. +

- Non-microarray related statistical/mathematical functions are available through the SAS® modules. +

- Extensible to new functions using XML. +

User Defined
- Pre-processing and transformation:
  - A number of algorithms are incorporated into the Microarray Solution:
  - Several state-of-the-art methods are missing, e.g., RMA, GC-RMA, MAS 5.0
  - User-defined transformations can be performed using DATA STEP and PROC TRANSPOSE, which are available as 'Utilities'.
Group comparisons:

- Available via the incorporated 'Mixed Model Analysis'.
- Several methods for multiple testing adjustment are offered within the 'Mixed Model Analysis':

- Other available SAS® procedures can be incorporated using XML.
SAS® Microarray Solution

- Group comparisons:
  - Permutation based testing
    - widely used in microarray settings.
    - not only in standard test-settings.
    - used to assess interestingness of any parameter (e.g., correlation,...).
    - is not yet supported directly, would need to be programmed.
SAS® Microarray Solution

• Clustering and Classification:
  ➔ Hierarchical Clustering incorporated.
  ➔ Principal Components Analysis incorporated.
  ➔ Linear Discriminant Analysis available.
SAS® Microarray Solution

- Clustering and Classification:
  - More recent methods not contained.
    - Diagonal linear discriminant analysis.
    - Support vector machines.
  - No general support for classifier assessment.
    - Split of data set into training- and validation-set.
    - Leave-k-out cross-validation.
    - Application of a generated classifier to a validation-set.
Annotation:

- Gene Ontology (The Gene Ontology Consortium (2000)) classification can be imported based on chip type.
- Mapping of Affymetrix names to Locus Link identifier is supported.
- Fisher's exact test is available for identification of interesting groups based on annotation classes.
- Direct link to different web sources from JMP.
Visualisation:

- Visualisation is done via JMP-interface.
  - Volcano plots and diagnostic plots for group comparisons.
  - Diagnostic plots for classification procedures.
  - PCA plots.

- Other graphs can be assembled via incorporation of user-defined procedures:
  - Histograms.
  - Profile plots.
  - Venn diagrams.
SAS® Genetic Marker Solution

- Statistical Genetics (Association Studies):
  - Test for Hardy-Weinberg Equilibrium and other marker properties available in 'Marker Properties'. +
  - Genotype analysis procedures available in 'Case-Control Association', tables are only presented via frequencies. +/-
  - Haplotype estimation possible via EM-algorithm. +
  - 'Haplotype Trend Regression' enables analysis of quantitative traits. +
Conclusions

- **Major strengths:**
  - Very good working tool for the SAS®-accustomed statistician.
  - Established methodology is available.
  - Sophisticated statistical functions in form of SAS/BASE® and SAS/STAT®, easy incorporation into analysis workflow.

- **Major drawbacks:**
  - Lack of most widely used preprocessing algorithms.
  - Caveat: R!
SDS at Schering AG

- SAS® Microarray Solution at Schering AG:
  - Clinical data is available in SAS® format.
  - Strong SAS® programming skills inhouse.
  - Ongoing discussion with SAS® Germany about further development of the tool.

- ... and a surprise:
  - SAS® Scientific Discovery Solution is not available with SAS® 9.1!
References

- FDA (2005): "Guidance for Industry: Pharmacogenomic Data Submission"
Contact

Susanne Schwenke
Schering AG
Medical Data Sciences, Non-Clinical Statistics
13342 Berlin
Germany

susanne.schwenke@schering.de