



ulm university universität  
**uulm**

HUMBOLDT-UNIVERSITÄT ZU BERLIN



**Exposé der Diplomarbeit:**

# Visualisierung von Genexpressionsdaten

Eingereicht von:

**Johannes Kozakiewicz**  
Institut für Informatik  
Humboldt-Universität zu Berlin  
Matr.Nr.: 186778  
[kozakiewicz@gmx.de](mailto:kozakiewicz@gmx.de)

Betreuer:

**Herr Dr. Hans Kestler**  
Institut für Neuroinformatik  
Universität Ulm  
AG Bioinformatik & Systembiologie  
[hans.kestler@uni-ulm.de](mailto:hans.kestler@uni-ulm.de)

Betreuer:

**Herr Prof. Dr. Ulf Leser**  
Institut für Informatik  
Humboldt-Universität zu Berlin  
Knowledge Management in Bioinformatics  
[leser@informatik.hu-berlin.de](mailto:leser@informatik.hu-berlin.de)

Ulm, den 20.10.2008

## 1. Motivation

Ein großer Bereich der Bioinformatik beschäftigt sich mit der Auswertung und Analyse von Microarray Experimenten. Microarray ist eine Sammelbezeichnung für moderne molekularbiologische Untersuchungssysteme, die die parallele Analyse von mehreren tausend Genen in geringen Mengen biologischer Proben erlauben. Es gibt verschiedene Formen von Microarrays, wie z.B. ArrayCGH, SNP und miRNA, die auch als "Genchips" oder "Biochips" bezeichnet werden, da sie wie ein Computerchip viele Informationen auf kleinstem Raum enthalten können. Mit ihrer Hilfe kann die Transkriptionsaktivität einer Zelle und somit die Aktivität bestimmter Gene zu einem bestimmten Zeitpunkt quantitativ gemessen werden. Zum Beispiel kann die Aktivität von Genen in einem gesunden und einem kranken Gewebe verglichen werden, um so möglicherweise die Ursachen für Erkrankungen festzustellen [1,9].

Die Daten, die bei einem solchen Experiment anfallen, sind von hoher Dimensionalität (z.B. Genanzahl) und geringer Kardinalität (Anzahl der Proben). Ohne entsprechendes Hintergrundwissen aus dem biologischen Fachbereich ist es bei solchen Datensätzen schwer, gefundene relevante Ergebnisse von zufälligen Ergebnissen, die durch Rauschen oder Messfehler entstehen können, zu unterscheiden. Deshalb werden neben der maschinellen Auswertung eines Experiments die Ergebnisse auch immer von einem menschlichen Experten begutachtet. Bei der Menge der Daten ist dies in der ursprünglichen tabellarischen Form eine schwierige, wenn nicht sogar unmögliche Aufgabe. Die Daten werden deswegen oft zuvor in eine grafische Darstellung gebracht, einem sogenannten Visualisierungsschema. Aufgrund der Anzahl der verschiedenen Eigenschaften werden schematische Darstellungen dieser Ergebnisse in univariate, bivariate oder multivariate Schemata unterschieden [2]. Mögliche Visualisierungsschemata für solche Daten wären Heatmaps oder Profile Plots [3]. Visualisierungsschemata sollen dem Betrachter dabei einen intuitiven Überblick über den Datensatz verschaffen und die Sachverhalte möglichst korrekt darstellen, sind dabei jedoch immer mit einem Informationsverlust verbunden [4,8].

### 2. Aufgabenstellung

In der Diplomarbeit sollen multivariate Visualisierungsschemata entworfen und evaluiert werden, welche Daten aus verschiedenen Quellen für die Visualisierung berücksichtigt, z.B. Schwellwerte, die Benutzer anhand einer Heatmap ermitteln können, oder Clusterinformationen aus einem Dendrogramm. Ebenfalls sollen Genexpressionsdaten und Metainformationen aus anderen, öffentlich zugänglichen Datenbanken wie z.B. GO (Gene Ontology) [5], GEO (Gene Expression Omnibus) [6] oder KEGG (Kyoto Encyclopedia of Genes and Genomes) [7] benutzbar sein. Denkbar wäre ein Vergleich von vorhandenen Genexpressionsdaten mit eigenen Experimentdaten oder eine Umsortierung der Daten nach bestimmten Metainformationen, wie z.B. die Involvierung bestimmter Gene in speziellen Pathways.

Zur Evaluierung dieser Schemata sollen psychophysische Vorstudien mit anderen Visualisierungsschemata, wie z.B. einer Heatmap oder eines Dendrogramms, im Bezug auf Lesbarkeit und Erkennbarkeit bestimmter Sachverhalte durchgeführt werden. Zu diesem Zweck soll eine Testumgebung selbst erstellt werden, in der noch zu definierende Aufgaben durch ca. 20 Probanden zu lösen sind. Die Probanden werden sich aus freiwilligen Studenten der Universität Ulm zusammensetzen. Das Testsetting soll als interaktiver Test implementiert werden, der anschließend maschinell mithilfe statistischer Methoden ausgewertet wird.

Aufbauend auf den Resultaten soll anschließend eine Java Applikation entwickelt werden, die multivariate Microarray-Daten mithilfe der entworfenen Schemata sowie anderer Schemata visualisiert. Merkmale aus einem Datensatz sollen dabei selektierbar sein und nach bestimmten Kriterien umsortiert werden können. Obligatorisch für die Applikation soll außerdem eine Schnittstelle zur Programmiersprache R sein, so dass auch andere Berechnungen durchgeführt und visualisiert werden können.

Optional wäre die Anbindung der Java Applikation an eine Datenbank, um Genexpressionsdaten sowie Resultate von Berechnungen zu laden und zu speichern. Neben dem relativ unabhängigen Datenformat und der Möglichkeit komplexe Berechnungsergebnisse zu speichern, wäre auch die simultane Bearbeitung und Visualisierung mehrerer Datensätze denkbar.

### 3. Vorgehen

Als erstes erfolgt der Entwurf mehrerer multivariater Schemata zur Visualisierung von Genexpressionsdaten. Der Entwurf wird auf grundlegende Visualisierungsverfahren wie der Visualisierung 2-dimensionaler Kohonen-Karten, multidimensionale Skalierung auf 2 bzw. 3 Dimensionen und Parallelkoordinaten-Plots aufbauen [8]. Welche Informationen durch welches Schema am besten visualisiert werden kann, soll im Verlauf der Arbeit geklärt werden. Nach dem Entwurf werden die Schemata mithilfe eines experimentellen Setups getestet. Durchgeführt werden die Experimente unter Aufsicht von Mitarbeitern der AG Bioinformatik & Systembiologie der Universität Ulm in einem Psychophysik-Labor des Instituts für Neuroinformatik. Inhalt und Ablauf des Setups werden noch im Verlauf der Diplomarbeit definiert. Im Allgemeinen wurden visuelle Wahrnehmungstests im Bezug auf den zu gewinnenden Informationsgehalt aus den Darstellungen vorgeschlagen.

Abhängig vom Evaluierungsergebnis wird dann entweder die Schemata einer weiteren Entwurfsphase unterzogen, um sie zu verbessern, oder in eine Java-Applikation vollständig implementiert. Anschließend werden die in der Aufgabenstellung beschriebenen Komponenten in die Applikation implementiert. Der Entwurf einer einfachen Datenbank und die Anbindung an die Applikation ist optional und soll nur bei entsprechend verfügbarer Zeit umgesetzt werden. Im letzten Schritt soll die Applikation ebenfalls auf Benutzerfreundlichkeit und Übersichtlichkeit getestet werden. Eventuell können hier weitere Verbesserungsmöglichkeiten bei der Bedienung und Visualisierung gefunden werden, die dann ebenfalls implementiert werden sollen.

### 4. Umsetzung

Die für die Implementierung und Evaluierung benötigten Genexpressionsdaten können entweder über bestimmte Paket-Bibliotheken von R geladen oder durch verschiedene Anbieter im Internet frei bezogen werden. Metainformationen, die bei der Analyse der Datensätze hilfreich sein können, werden durch öffentlich, zugängliche Datenbanken wie KEGG oder GO bezogen. Diese Informationen werden in die Applikation integriert, so dass sie für selektive Aufgaben genutzt werden können.

Die Implementierung der Schemata und der Applikation erfolgt in Java. Berechnungen zu Clusterinformationen und anderen statistisch relevanten Informationen werden durch in R-Pakete bereitgestellte Funktionen berechnet und anschließend an die Applikation übergeben.

## Referenzliste

- [1] Patrick O. Brown, David Botstein: **Exploring the new world of the genome with DNA microarrays**. Departments of Biochemistry and Genetics and the Howard Hughes Medical Institute, Stanford University School of Medicine, California, 1999  
(URL: [http://www.nature.com/ng/journal/v21/n1s/pdf/ng0199supp\\_33.pdf](http://www.nature.com/ng/journal/v21/n1s/pdf/ng0199supp_33.pdf))
- [2] Hans-Friedrich Eckey, Reinhold Kosfeld, Martina Rengers: **Multivariate Statistik**. Gabler Verlag, Wiesbaden, 2002
- [3] Wei Yin, Qing Yang, Shaoyuan Yao, Yong Shi: **Hierarchical Clustering Analysis from Genomic Dataset**. Intelligent Agent Technology, IEEE/WIC/ACM International Conference on 19-22 Sept. 2005, Compiègne University of Technology France, Page(s): 759- 762, Digital Object Identifier 10.1109/IAT.2005.81  
(URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01565638>)
- [4] Nils Gehlenborg: **Visualisierung für DNA-Chip-Analysen**. Center for Bioinformatics Tuebingen (ZBIT), University of Tuebingen, 2005  
(URL: [www.zbit.uni-tuebingen.de/pas/mayday/preprints/nils\\_gehlenborg\\_doit2005.pdf](http://www.zbit.uni-tuebingen.de/pas/mayday/preprints/nils_gehlenborg_doit2005.pdf))
- [5] Gene Ontology, URL: <http://www.geneontology.org/>, Oct. 2008
- [6] Gene Expression Omnibus, URL: <http://www.ncbi.nlm.nih.gov/geo/>, Oct. 2008
- [7] Kyoto Encyclopedia of Genes and Genomes, URL: <http://www.genome.jp/kegg/>, Oct. 2008.
- [8] Colin Ware: **Information Visualization: Perception für Design**. 2nd Edition, 2004, Elsevier, San Francisco
- [9] Markus Weniger: **Genome Expression Pathway Analysis Tool - Analyse und Visualisierung von Microarray Genexpressionsdaten unter genomischen, proteomischen und metabolischen Gesichtspunkten**. Dissertation Julius-Maximilians-Universität Würzburg, 2007