Performanzvergleich von Genexpressionsdatenbanken

Exposé

Muhammad Ghiyas

Betreuer: Prof. Dr. Ulf Leser
Institut für Informatik, Humboldt-Universität zu Berlin

Zeitraum: 15.02.04-14.08.04

Motivation

Durch Genexpressionsanalyse werden Eigenschaften der Gene und ihre Einflüsse auf den lebenden Organismus charakterisiert. Es gibt viele kommerzielle und zu Forschungszwecken erstellte Datenbanken, einige davon sind öffentlich zugänglich. Zum Beispiel ArrayExpress [1], GeneX [2], MCHIPS [3], SMD [4] etc. In molekularbiologischen Forschungen spielen Genexpressionsexperimente eine große Rolle. Bisher wurden die Daten von Experimenten in einem Flatfile gespeichert. Durch die Anwendung von relationalen Mikroarray Datenbanken ist es leichter geworden, Genexpressionensdaten zu speichern und zu analysieren. Eine Mikroarray Datenbank ermöglicht es die Ergebnisse von tausenden Gene in vielen Experimenten gleichzeitig zu verarbeiten. Da die Datenmenge immer größer wird, werden Datenbanken eine immer größere Rolle spielen. Zunehmend wird man Datenbanken auch deswegen benutzten, weil die Datenmengen so groß werden, dass ihre Analyse nicht mehr im Hauptspeicher erfolgen kann.

Bis jetzt sind Genexpressionsdatenbanken nur unter qualitativen Gesichtspunkten verglichen worden. Es fehlt noch ein quantitativer Performanzvergleich solcher Datenbanken.

Zielsetzung

Die Aufgabe dieser Diplomarbeit ist es, einen Performanzvergleich einiger öffentlicher Genexpressionsdatenbanken durchzuführen. Dazu werden vier Genexpressionsdatenbanken lokal implementiert. Der Performanzvergleich der Datenbanken wird anhand typischer Analysen von Genexpressionsdaten gemacht. Typische Analysen von Genexpressionsdaten sind folgende:

- Differentielle Analyse: Bei welchen Genen verändern sich unter veränderten Bedingungen die Expressionen
- Ko-Regulation von Genen: Welche Gene reagieren gleich auf Veränderungen in der experimentellen Umgebung

Vorgehen

Die Diplomarbeit baut auf Vorarbeiten auf, die in den letzten Monaten am Lehrstuhl „Wissensmanagement in der Bioinformatik, Humboldt Universität zu Berlin“ gemacht wurde.
Die drei Datenbanken ArrayExpress, GeneX und MCHIPS sind schon vorhanden. Die Diplomarbeit wird wie folgt aufgebaut sein:

1 Untersuchung von Genexpressionsdatenbanken
2 Implementierung der ausgewählten Genexpressionsdatenbanken
3 Performanzvergleich nach verschiedenen Kriterien
4 Optimierung und Verbesserungsvorschläge

Untersuchung und Implementierung von Genexpressionsdatenbanken
Die Genexpressionsdatenbanken werden untersucht. Für diese Untersuchung werden folgende Datenbanken ausgewählt. ArrayExpress [1], GeneX [2], MCHIPS [3] und SMD [4].
Die ausgewählten Datenbanken werden auf dem RDBMS ORACLE (Version 9.2) implementiert. Die Datenbanken werden mit gleichen Daten [5] gefüllt. Durch Parser werden die Daten aus dem Flatfile gelesen und in die Datenbanken geschrieben.

Performanzvergleich
Die Diplomarbeit besteht aus typischen Anfragen, nämlich der differenziellen Analyse und Ko-Regulation von Genen. Vorausgehend wird die Varianz pro Gen wird berechnet. Das kann später zum Filtern benutzt werden, d.h. nur Gene mit einer bestimmten Varianz werden in die weiteren Berechnungen mit einbezogen.
Zur differentiellen Analyse müssen die Expressionsdaten in Gruppen eingeteilt werden. Im Falle der von uns benutzten Testdaten unterscheiden sich diese Gruppen in der spezifischen Art von Leukämie, an der Patient, dessen Genexpression gemessen wurde, erkrankt war.
Die differenzielle Analyse wird durch zwei Methoden berechnet:
1. T-Statistik
2. SAM
Zur Bestimmung von Ko-Regulation wird ein hierarchischer Clusteralgorithmus benutzt. Die Clusterrepräsentanten werden als Mittelpunkt aller Clusterelemente berechnet.
Diese Verfahren werden in der Datenbank auf den vier verschiedenen Schemata implementiert. Die Daten werden in verschiedenen Szenarien untersucht und die Ergebnisse werden wieder in einer Datenbank gespeichert. Durch Vergleich der Ergebnisse wird die Korrektheit der Implementierung sichergestellt (dadurch werden verschiedene Implementierungen identischer Algorithmen auf unterschiedlichen Schemata mit gleichen Daten vergleichen). Die Analysezeiten werden notiert und verglichen. Für ausgewählte Schemata wird außerdem eine Optimierung vorgeschlagen und deren Erfolg anhand weiterer Messungen nachgewiesen.

Referenzen
[1] Brazma, A., Parkinson, H.,Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., et al.(2003).
“ArrayExpress-- a public repository for microarray gene expression data at the EBI.”
Nucleic Acids Res 31(1): 68-71.

[2] Mangalam, H., Stewart, J., Zhou, J., Schlauch, K., Waugh, M., Chen, G., Farmer, A. D., Colello, G. and Weller, J. W. (2001).
“GeneX: An Open Source Gene Expression Databases and Integrated Tool Set.”
IBM Systems Journal 40(2): 552-569.

[3] Fellenberg, K., Hauser, N. C., Brors, B., Hoheisel, J. D. And Vingron, M. (2002).
“Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis.”
Bioinformatics 18(3): 423-433.

[4] Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, JC., Dwight, SS., Kaloper, M., Weng, S., Jin, H., Ball, CA., Eisen, MB., Spellman, PT., Brown, PO., Botstein, D., Cherry, JM. (2001)
„The Stanford Microarray Database.“
Nucleic Acids Res 1;29(1):152-5.

[5] Daten

ALL1:
Higuchi, M., O'Brien, D., Kumaravelu, P., Lenny, N., Yeoh, EJ., Downing, JR. (2002).
„Expression of a conditional AML1-ETO oncogene bypasses embryonic lethality and establishes a murine model of human t(8;21) acute myeloid leukemia“.
Cancer Cell.;1(1):63-74.
ALL3:
Ross, E. M., Zhou, X., Song, G., Shurtleff, A. S., Girtman, K., Williams, K. W., Liu, H., Mahfouz, R., Raimondi, C. S., Lenny, N., Patel, A., Downing, R. J. (2003).
„Classification of pediatric acute lymphoblastic leukemia by gene expression profiling“.
Blood 102: 2951-2959.
(St. Jude Children's Research Hospital) http://www.stjuderesearch.org/

[6] T-test
Lozan, L. J., Hartmut, K. (1998).
“Angewandte Statistik für Naturwissenschaften.”
2. Auflage, Seite 110-112

[7] SAM
Tusher, VG., Tibshirani, R., Chu, G. (2001).
“Significance analysis of microarrays applied to the ionizing radiation response.”
Proc Natl Acad Sci U S A. 24;98(9):5116-21.
http://www-stat.stanford.edu/~tibs/SAM/pnassam.pdf