Boehringer Ingelheim

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Exposé zur Diplomarbeit

ANALYSE VON GENEXPRESSIONSDATEN FÜR DIE ERFORSCHUNG VON ARZNEISTOFFEN

Eingereicht von:

Johannes Kozakiewicz
Institut für Informatik
Humboldt-Universität zu Berlin
Matr.Nr.: 186778
Email: kozakiewicz@gmx.de

Betreuung:

Prof. Dr. Ulf Leser
Institut für Informatik, Humboldt-Universität zu Berlin
Email: leser@informatik.hu-berlin.de
Tel.030-2093-3902

Dr. Eric Simon
Boehringer Ingelheim, Pharma GmbH & Co. KG
Research, Pulmonary Diseases Research
Email: eric.simon@boehringer-ingelheim.com
Tel. 07531-54 94883

Kooperation:

Boehringer Ingelheim Pharma GmbH & Co. KG und die Humboldt-Universität zu Berlin

Boehringer Ingelheim

Einleitung

Forschende Pharmaunternehmen wie Boehringer-Ingelheim bedienen sich zunehmend neuerer molekularbiologischer Methoden um innovative Arzneimittel zu finden und zu entwickeln. Eine dieser Methoden ist die Verwendung von DNA Microarrays zur genomweiten Bestimmung der RNA Expression biologischer Gewebe- oder Zellproben [1]. Ein großer Bereich der Bioinformatik beschäftigt sich mit der Auswertung und Analyse solcher Genexpressions-Experimente. Die Daten, die bei einem solchen Experiment anfallen, sind von hoher Dimensionalität und geringer Kardinalität. Das heißt, einer vergleichsweise geringen Anzahl an gleichzeitig erfassten Proben steht eine hohe Anzahl an beobachteten Genen gegenüber. Ziel ist es unter anderem durch vergleichende Analyse unterschiedlicher Proben (z.B. gesund und erkrankt) potentielle Angriffspunkte für neue Arzneiwirkstoffe zu identifizieren [2,3]. Die Ergebnisse solcher Analysen sind nicht immer leicht zu interpretieren, da die zugrunde liegenden Messungen fehlerbehaftet sind und die biologischen Proben vielen Einflussfaktoren unterliegen. Ohne biologisches Hintergrundwissen ist es deshalb schwer, biologisch relevante Ergebnisse von Rauschen, Messfehlern oder experimentellen Artefakten zu unterscheiden.

Öffentliche Datenquellen gewinnen zunehmend an Bedeutung für die Arzneimittelforschung. Verbesserte Technologien, die wachsende Zahl und Größe wissenschaftlicher Projekte sowie die generelle Forcierung der Offenlegung von Forschungsresultaten durch die Forschungspolitik und Zeitschriftenverlage fördern die Bereitstellung solcher Datenquellen. Aufgrund der großen Anzahl der mittlerweile zur Verfügung stehenden Experimente ist es dringend notwendig diese systematisch zu erschließen d.h. alle wichtigen Quellen zu kennen und relevante Experimente für ein Indikationsgebiet mit Hilfe eines automatisierten Verfahrens vorzufiltern. Mögliche Ansatzpunkte, um diesen Prozess zu unterstützen, sind der Einsatz von Ontologien und Text Mining Verfahren. Die Komplexität der Analyse der Datensätze steigt dabei mit der Anzahl der ausgewählten Datensätze, aber auch mit der Zahl der unterschiedlichen Datenquellen, Microarray-Technologien und Tierspezies sowie den verschiedenen Indikationen[1]. Die Herausforderung wird darin bestehen, die existierenden Verfahren wie die sogenannte Gene Set Enrichment Analysis [3] oder Gene Expression Network Analysis [4] und klassische differentielle Expressionsanalyse [7], in sinnvoller Weise miteinander zu kombinieren, um neue interessante Zielgene für die untersuchten Krankheiten zu finden.

1 Anwendungsgebiet, Heilanzeige

Ziel

Ziel der Diplomarbeit ist die Erschließung, Aufbereitung und Analyse öffentlicher Quellen genomweiter Expressionsdaten mit Bezug zu Erkrankungen des zentralen Nervensystems (Parkinson, Alzheimer, Chronischer Schmerz), der Lunge (Asthma und COPD[2]) und des Stoffwechsels (Arteriosklerose[3], Adipositas[4], Typ II Diabetes).

Unter den existierenden Chiptechnologien soll zunächst für Mensch, Maus und Ratte auf Affymetrix Genome Arrays (HG-U133 der Subtypen A, B sowie Plus_2, Mouse_430_2, und Rat_230_2) [L1] und Illumina Bead Chips (HumanHT-12_V3, HumanRef-8_V2, HumanRef-8_V3, MouseRef-8_V1_1, MouseRef-8_V2 und RatRef-12_V1) [L2] zurückgegriffen werden. Entsprechende Annotationen sind vorhanden. Als Datenquellen dienen die beiden großen Repositorien für öffentliche Daten dieser Art, Gene Expression Omnibus (GEO) [5] am National Center for Biotechnology Information (NCBI) [L3] und ArrayExpress [6] am European Bioinformatics Institute (EBI) [L4]. Darüber hinaus gibt es eine Reihe weiterer Quellen, deren Erschließung ein Teil der Aufgabenstellung darstellt. Beispielhaft seien hier genannt das Diabetes Genome Anatomy Project [8, L9] zu Diabetes sowie die Connectivity Map [9] am Broad Institute [L10] mit über 7000 Expressionsprofilen von Zelllinien, die mit 1309 zugelassenen Arzneistoffen behandelt wurden.

In einem ersten Schritt werden für jede Indikation Schlagwortlisten entwickelt - in Zusammenarbeit mit biologischen Experten und unter Zuhilfenahme von vorhandenen Ansätzen [L5]. Anschließend werden die zur Verfügung stehenden Datensätze den einzelnen Indikationen zugeordnet. Diese automatische Klassifikation erfolgt durch Text Mining auf Basis der zur Verfügung stehenden Metainformationen (Experiment Annotationen, Publikationen). Sie soll – zumindest beispielhaft – ebenfalls in Zusammenarbeit mit biologischen Experten aus dem Indikationsgebiet überprüft und ergänzt werden..

Der inhaltliche Schwerpunkt der Arbeit soll in der anschließenden Analyse liegen.. Dazu werden anhand der zur Verfügung stehenden Datensätze eine oder zwei Indikationen ausgewählt und die entsprechenden Daten mithilfe von existierenden Verfahren analysiert (differentielle Expressionsanalyse [7], Gene Set Enrichment Analysis [3] und Gene Expression Network Analysis [4]) um interessante Ziel- und Markergene für die priorisierte(n) Indikation(en) zu finden. Tools, die diese Verfahren implementieren, existieren bereits, sodass hier der Fokus auf der Analyse der Ergebnisse liegt. Ziel dieser Analyse ist die immer zahlreicher werdenden,

2 Chronisch obstruktive Atemwegserkrankung (engl.) , 3 Arterienverkalkung , 4 Fettsucht

öffentlich zugänglichen Experimente mit in die eigenen Analysen einzubeziehen. Dabei kommt der vergleichenden Analyse verschiedener Experimente eine besondere Bedeutung zu, da sie die Konfidenz der Aussagen bezüglich einzelner Gene vervielfacht bzw. eine wechselseitige Validierung ermöglicht. Wenn zum Beispiel ein Transkript in einem Experiment signifikant dereguliert ist und diese Signifikanz durch ein oder mehrere weitere Experimente gestützt wird, dann erhöht dies die Wahrscheinlichkeit einer Relevanz des kodierten Proteins für den untersuchten Prozess.

Als zusätzliche unabhängige Validierung soll ein Ansatz zum Abgleich mit Literaturdaten aus öffentlichen Datenbanken wie PubMed [L8] verfolgt werden. Dazu werden die Treffer aus PubMed für die entwickelten Schlagwortlisten aus dem 1.Schritt sowie für die Namen der Zielgene mithilfe vorhandener Tools, wie z.B. Temis [L11], miteinander verglichen.

Vorgehensweise

1) Zusammenstellung der Datenquellen

- Quellen: Gene Expression Omnibus, Array Express, Diabetes Genome Anatomy, Connectivity Map, Journals, u.a.
- Übersicht zu verfügbaren Datensätzen, Technologien, Spezies, Dynamik (Updates)
- Übersicht zu vorhandenen Tools (API, Tabellenschemata, Webservice, Visualisierung)
- Metainformationen und Publikationen
- *Fachliche Unterstützung*: Bioinformatik Team

2) Zusammenstellen der Datensätze

- Festlegung der zu verwendenden Datenquellen, Chip-Technologien und Spezies
- Auswahl der zu verwendenden Datensätze
- *Fachliche Unterstützung*: Bioinformatik Team

3) Erstellung von Schlagwortlisten für jede Indikation

- Krankheiten (zum Teil vorhanden für Asthma)
- Zielgewebe (z.B. Asthma→Lunge)
- Bekannte Ziel- oder Markergene (optional)
- *Fachliche Unterstützung*: Bioinformatik Team, Indikation und wissenschaftliche Bibliothek

4) Klassifikation der Daten und Auswahl des Schwerpunkte

- Attributzuweisung: Indikation, Studie, Gruppe
- Eventuell zusätzliche Attribute (Anatomie, Behandlung)
- *Fachliche Unterstützung*: Bioinformatik Team und Indikation

Boehringer Ingelheim

5) Analyse

- Prozessierung der Daten (Workflow vorhanden)
- Differentielle Analyse, Enrichment Analyse, Netzwerk Analyse
- Vergleichende Analyse verschiedener Experimente
- Validierung von Kandidatengenen mit Hilfe von Text Mining
- *Fachliche Unterstützung*: Bioinformatik Team, Indikation und wissenschaftliche Bibliothek

6) Publikation

Literatur

1. Antonia Humm, Prof. Dr. Andreas Busch, Dr. Kemal Malik: Science for a better life – Vom Molekül zum Medikament: Exkursion in die Forschung, Bayer Schering Pharma AG, 2008.

2. 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

3. Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler & Leif C: Groop4, PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, 2003; 34: 267 - 273.

4. Nacu, S., R. Critchley-Thorne, P. Lee und S. Holmes: Gene expression network analysis and applications to immunology. Bioinformatics 2007; 23(7):850–858.

5. Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar, NCBI GEO: mining tens of millions of expression profiles—database and tools update, Nucleic Acids Res., 2007; 35: D760 - D765.

6. Helen Parkinson, Misha Kapushesky, Nikolay Kolesnikov, Gabriella Rustici, Mohammad Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Ele Holloway, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Tim F. Rayner, Faisal Rezwan, Anjan Sharma, Eleanor Williams, Xiangqun Zheng Bradley, Tomasz Adamusiak, Marco Brandizi, Tony Burdett, Richard Coulson, Maria Krestyaninova, Pavel Kurnosov, Eamonn Maguire, Sudeshna Guha Neogi, Philippe Rocca-Serra, Susanna-Assunta Sansone, Nataliya Sklyar, Mengyao Zhao, Ugis Sarkans, and Alvis Brazma;
ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res., January 2009; 37: D868 - D872.

7. Cui, X. und G. A. Churchill: Statistical tests for differential expression in cDNA microarray experiments. Genome Biology, 2003; 4(4)210.

8. Beckley, Elizabeth Thompson: Gene Gives New Insight Into Diabetes. DOC News 2006 3: pages: 1-11.

9. Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr, Eric S. Lander, and Todd R. Golub: The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease, Science vol.313 (number 5795), pages: 1929-1935; 29 September 2006.

10. Affymetrix, URL: http://www.affymetrix.com/, Apr. 2009

11. Illumina, URL: http://www.illumina.com/, Apr. 2009

12. Gene Expression Omnibus, URL: http://www.ncbi.nlm.nih.gov/geo/, Apr. 2009

13. ArrayExpress, URL: http://www.ebi.ac.uk/microarray-as/ae/, Apr. 2009

14. The Open Biomedical Ontologies, URL: http://www.obofoundry.org/, Apr. 2009

15. PubMed, URL: http://www.ncbi.nlm.nih.gov/pubmed/, Apr. 2009

16. Diabetes Genome Anatomy Project, URL: http://www.diabetesgenome.org, Apr. 2009

17. Broad Institute's Connectivity Map, URL: http://www.broad.mit.edu/node/305, Apr. 2009

18. Temis, URL: http://www.temis.com, Apr. 2009