Exposé zur Diplomarbeit

Integration und Scoring von Pathways

Oliver Arnold

Betreuer: Ulf Leser

1 Hintergrund

Ein zentrales Problem der Biologie ist die Identifikation und Zuordnung von Genprodukten, die an biologischen Prozessen beteiligt sind. Durch die Entwicklung von DNA-Microarrays ist es möglich geworden die Expressionslevel von Genen unter bestimmten Bedingungen zu ermitteln und physiologischen Zellstadien zuzuordnen. Expressionslevel geben an wie stark Genprodukte in der Zelle synthetisiert werden. Oft werden die Genexpressionslevel von zwei Zellstadien miteinander verglichen um die unterschiedlichen Auswirkungen von Einflüssen auf Zellen zu untersuchen [NCL+07]. Diese Methode wird unter anderem zur Entwicklung von Medikamenten oder neuer Behandlungsmethoden verwendet.

Allerdings werden durch die Verwendung von Microarrays viele Daten generiert, deren Analyse und Interpretation sich schwierig gestaltet [TFB07]. Um Genexpressionen nicht nur einzeln zu betrachten, sondern den Zusammenhang von mehreren Genexpressionen in einem biologischen Prozess zu verstehen, ist es sinnvoll Expressionen chemischen Komponenten in Pathways zuzuordnen. Pathways sind biologische Prozesse, die als Graph dargestellt werden. Es gibt unterschiedliche Arten von Pathways ([BCS06]):

- **metabolische Pathways:** Netzwerke von biochemischen Reaktionen, die an Umwandlungen von Metaboliten beteiligt sind
- **Signal-Pathways:** molekulare Interaktionen und chemische Modifikationen, durch die Zellen auf äußere Reize reagieren können
- **Genregulations-Netzwerke:** Transkriptions- und Translationsfaktoren, die die Expressionen von Genen regulieren und damit die Konzentration von Genprodukten in der Zelle steuern
- **Protein-Compound und Protein-Protein-Interaktionen:** geben jeweils die Wechselwirkung zwischen Proteinen oder Protein und Compound an

Es gibt viele Pathway-Datenbanken, die meistens einen Fokus auf unterschiedliche biologische Prozesse haben; z.B. die Entwicklung von Tumoren. Außerdem gibt es zur Zeit keine definierten Grenzen für Pathways, bzw. wie der gesamte Stoffwechselweg eines Organismus in einzelne Pathways unterteilt werden soll [GK06]. Zum Beispiel wird bei der Erstellung von Pathways in BioCyc versucht Regeln einzuhalten, die beschreiben was ein Pathway enthalten sollte ([GK06]):

1. Finden von einem gemeinsamen biologischen Prozess: eine Menge von Reaktionen, die fast zur gleichen Zeit stattfinden, unter den gleichen Bedingungen aktiv werden und sich nicht gegenseitig ausschließen
2. Pathwaygrenzen bei Substraten, die in vielen Reaktionen verwendet werden

3. Pathwaygrenzen bei stabilen Substraten (bleiben über einen längeren Zeitraum erhalten), nicht bei flüchtigen
4. einzelne Schritte in Pathways sollen einer gemeinsamen Regulation unterliegen
5. Pathways sind evolutionär konserviert

Bei KEGG hingegen sind Pathways aus Segmenten von verschieden Organismen zusammengesetzt und enthalten Abläufe von Synthese und Degradation eines Metabolits. Durch die Kombination von mehreren biologischen Prozessen und alternativen Wegen für Substrate sind Pathways der KEGG-DB daher meistens größer. Durch die unterschiedlichen Verfahren zur Erstellung von Pathways kommt es zu Überschneidungen zwischen Pathways von verschiedenen Pathway-Datenbanken.

2 Motivation

In der vorangegangen Studienarbeit wurden bereits Genexpressionen chemischen Komponenten in Pathways zugeordnet. Allerdings wurden dort nur Pathways aus der KEGG-DB zur Darstellung verwendet. Des Weiteren konnte nur ein Teil der aus einem Experiment zur Verfügung stehenden Genexpressionen Pathways zugeordnet werden. Das lag einerseits daran, dass nur ein Mapping zwischen Genen und Enzymen stattfand und Mappings zwischen Genen und anderen Proteinen nicht berücksichtigt wurden. Andererseits enthielten die Pathways der KEGG-DB nur einen Teil aller Enzyme, denen bei dem Experiment Genexpressionen zugeordnet wurden. Um einen möglichst großen Teil von Genexpressionen Pathways zuordnen zu können, sollen in dieser Arbeit möglichst viele Pathways durch verschiedene Datenbanken abgedeckt werden.

Der Benutzer des zu entwickelnden Programms soll ein Überblick darauf bekommen, welche Pathways durch die veränderten Genexpressionen beeinflusst wurden. Dazu wurde bisher nur ein einfacher Algorithmus benutzt. In dieser Arbeit soll unter anderem ein neuer Ansatz zur Einschätzung der Beeinflussung von Pathways durch Genexpressionen untersucht werden, der sich Verbindungen zwischen verschiedenen Pathways zunutze macht. Zwar wird bei den meisten Pathways angegeben, in welchen anderen Pathways die Produkte von Reaktionen weiter verwendet werden. Es könnten aber auch Moleküle (Proteine, Enzyme, Metabolite, etc.) für die Einschätzung der Auswirkung von veränderten Genexpressionen auf Pathways verwendet werden, für die keine Angabe darüber existieren in welchen Pathways sie an Reaktionen teilnehmen.

Um die Auswirkungen von veränderten Genexpressionen auf Signalpathways zu bewerten, sollte die Position des betroffenen Genprodukts berücksichtigt werden. Desto weiter am Anfang ein Genprodukt in der Signalkette beteiligt ist, umso stärker sollte die Auswirkung auf ein Signalpathway eingestuft werden.

3 Ziele

Verbindungen zwischen Pathways - durch Moleküle, die in mehreren Pathways an Reaktionen beteiligt sind - sollen zur Einschätzung der Auswirkung von Genexpressionen auf Pathways verwendet werden. Ein Molekül, das durch Genexpressionen in einem Pathway beeinflusst wurde, sollte wiederum Reaktionen in anderen Pathways beeinflussen, in denen das Molekül weiterverwendet wird.

Des Weiteren sollen Überscheidungen, bzw. isomorphe Subgraphen, zwischen Pathways verschiedener Datenquellen bestimmt werden. Pathways sollen in einer Liste anhand der Anzahl der Überschneidungen mit anderen Pathways sortiert aufgelistet werden. Außerdem soll bei der Selektion eines Pathways zur Visualisierung eine weitere Liste angegeben werden, die alle Pathways auflistet, die Überschneidungen mit dem aktuellen Pathway haben. Bei der Auswahl eines Pathways aus dieser Liste sollen Subgraphen, die in dem ausgewählten Pathway vorkommen, im aktuellen Pathway farblich hervorgehoben werden.

Es sollen verschiedene statistische Methoden zur Einschätzung der Signifikanz von Genexpressionen und der Einstufung der Auswirkung von veränderten Genexpressionen auf Pathways eingesetzt werden.

Alle vom Programm verwendeten Daten sollen in einer lokalen Datenbank gespeichert werden.

4 Vorgehen

Um die verschiedenen Pathways auf isomorphe Subgraphen untersuchen zu können, müssen alle Pathways in einem einheitlichen, vergleichbaren Format gespeichert werden. Da es für Proteine und andere chemische Komponenten mehrere Schreibweisen gibt, müssen diese mittels einer Datenbank verglichen werden. Zum Beispiel werden in Entrez Gene[1] mehrere Namen für ein Genprodukt aufgeführt; auch BRENDA[2] wäre geeignet, allerdings enthält diese Datenbank nur Informationen zu Enzymen.

mögliche Pathway-Datenbanken:

MetaCyc [3]: enthält mehr als 1100 Pathways von 1500 verschiedenen Organismen

KEGG [4]: enthält viele metabolische und Signal-Pathways; überwiegend metabolische

Reactome [5]: enthält viele regulatorische, metabolische und Signal-Pathways

Pathway Interaction Database [6]: enthält Signal-Pathways; zur Zeit 79 eigene und 317 Pathways, die von BioCarta und Reactome importiert wurden

[1] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
[2] http://www.brenda-enzymes.info
[3] http://www.metacyc.org
[4] http://www.genome.jp/kegg
[5] http://www.reactome.org
[6] http://pid.nci.nih.gov

PathwayCommons [7]: enthält über 1100 Pathway aus verschiedenen Datenbanken (Cancer Cell Map, HRPD, HumanCyc, IntAct, MINT, Pathway Interaction DB und Reactome)

InnateDB [8]: Fokus auf Immun-bezogene Pathways; importiert allerdings Pathways von vielen anderen Datenbanken

INOH [9]: enthält Signalpathways

SPIKE [10]: enthält Signalpathways

Bis auf KEGG, Reactome und InnateDB stellen alle Datenbanken die Pathways im BioPax-Format zum Download bereit [SL05]. KEGG verwendet ein eigenes Format KGML, Reactome stellt unter anderem eine Web Service API bereit und InnateDB benutzt das PSI-MI Format.

Anschließend können Pathways miteinander verglichen werden. Dieser Vergleich kann auf das Finden des maximalen gemeinsamen Subgraphen (MCS) zurückgeführt werden. Allerdings ist zu beachten, dass die Anzahl der möglichen Knoten im Vergleichsgraphen, auf die der aktuelle Knoten gemappt werden kann, durch die Label stark eingeschränkt ist und das Problem des MCS dadurch stark vereinfacht wird. Beim Vergleich von Pathways ist zu beachten, dass ein Genprodukt mehrmals in einem Pathway verwendet werden kann. Daher wird ein Knoten durch sein Label nicht eindeutig identifiziert.

Um Moleküle zur Bewertung von anderen Pathways zu verwenden, müssen zuerst Genexpressionen bestimmt werden, die signifikant verändert sind. Dazu können verschiedene statistische Tests verwendet werden (siehe z.B. [KLS06]). Diese Tests geben jedoch kein Wert zurück, mit dem man Genexpressionen als hoch- oder runterreguliert einstufen kann. Daher werden für alle signifikanten Genexpressionen die Veränderung gegenüber dem Kontrollexperiment mit der Fold-Change Methode ([CC03]) berechnet.

Anschließend können die Moleküle, die durch veränderte Genexpressionen beeinflusst sind, zur Bildung von Verbindungen zwischen Pathways verwendet werden. Damit nicht zu viele Verbindungen entstehen, sollten die am häufigsten vorkommenden Moleküle nicht benutzt werden; alternativ wird eine Scoring-Methode entwickelt, die den Einfluss von Molekülen auf einen bestimmten Pathway auch von der Spezifität des jeweiligen Moleküls abhängig macht [Rob04]. Zur Ermittlung der Häufigkeit des Vorkommens von Molekülen in Pathways, können einfach alle importierten Pathways durchsucht werden.

Durch die Verbindungen zwischen Pathways entsteht ein gerichteter Graph, der auch Zyklen enthalten kann. Für jeden Pathway können jetzt die eingehenden Kanten ermittelt werden, wobei nur die Kanten beachtet werden, die als Startknoten - in einem anderen Pathway - ein signifikant beeinflusstes Molekül besitzen. Über die ermittelten Kanten sollen die vorher berechneten Werte (prozentuale Veränderungen) der Startknoten an Nachbarknoten im aktuellen Pathway propagiert werden. Im Paper [MI07] wird die Ausbreitung von Konzentrationsveränderungen von Proteinen in Protein-Protein-Interaktions Netzwerken untersucht. Es wurde festgestellt, dass die Auswirkung auf Nachbarproteine exponentiell mit der Entfernung zwischen Proteinen abnimmt und signifikante Auswirkungen in den meisten Fällen nur bei

[7] http://www.pathwaycommons.org/pc
[8] http://www.innatedb.ca
[9] http://www.inoh.org
[10] http://www.cs.tau.ac.il/~spike

Nachbarproteinen zu messen sind, die nicht mehr als zwei Interaktionen voneinander entfernt sind.

Danach können die Moleküle, die durch Moleküle aus anderen Pathways beeinflusst werden, erneut auf signifikante Veränderungen überprüft bzw. als signifikant beeinflusst eingestuft werden, sofern ihnen keine Genexpressionen zugeordnet wurden. Anschließend kann der aktuelle Pathway bewertet werden, z.B. mittels dem exakten Fisher-Test ([DKT+07]). Für die Auswirkung von Konzentrationsveränderungen von Molekülen auf andere Pathways werden die aus den Genexpressionsdaten berechneten Werte verwendet.

Da viele Pathway-Datenbanken nur die Beziehung (Kanten) zwischen chemischen Komponenten (Knoten) eines Pathways angeben, aber keine Positionsangaben, wird Prefuse zur Berechnung des Layouts für diese Pathways verwendet.

Die Gene Expression Omnibus ([BTW+07]) Datenbank beinhaltet viele Genexpressionsdatensätze, die zum Mapping auf Pathways verwendet werden können.

Literatur

[BCS06] Bader, G. D., M. P. Cary und C. Sander: *Pathguide: a Pathway Resource List.* Nucleic Acids Research, 34:D504–D506, 2006. 1

[BTW+07] Barrett, T., D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky und R. Edgar: *NCBI GEO: mining tens of millions of expression profiles–database and tools update.* Nucleic Acids Research, 35(Database issue):760–765, 2007. 5

[CC03] Cui, X. und G. A. Churchill: *Statistical tests for differential expression in cDNA microarray experiments.* Genome Biol, 4(4):210, 2003. 4

[DKT+07] Draghici, S., P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu und R. Romero: *A systems biology approach for pathway level analysis.* Genome Research, 17(10):1537–1545, 2007. 5

[GK06] Green, M. L. und P. D. Karp: *The outcomes of pathway database computations depend on pathway ontology.* Nucleic Acids Research, 34(13):3687, 2006. 1

[KLS06] Kim, S. Y., J. W. Lee und I. S. Sohn: *Comparison of various statistical methods for identifying differential gene expression in replicated microarray data.* Statistical Methods in Medical Research, 15(1):3–20, 2006. 4

[MI07] Maslov, S. und I. Ispolatov: *Propagation of large concentration changes in reversible protein-binding networks.* Proceedings of the National Academy of Sciences, 104(34):13655–13660, 2007. 4

[NCL+07] Nacu, S., R. Critchley-Thorne, P. Lee und S. Holmes: *Gene expression network analysis and applications to immunology.* Bioinformatics, 23(7):850–858, 2007. 1

[Rob04] Robertson, S.: *Understanding inverse document frequency: on theoretical arguments for IDF.* Journal of Documentation, 60(5):503–520, 2004. 4

[SL05] Stromback, L. und P. Lambrix: *Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX.* Bioinformatics, 21(24):4401–4407, 2005. 4

[TFB07] Trevino, V., F. Falciani und H. A. Barrera-Saldaña: *DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research.* Molecular Medicine, 13(9-10):527, 2007. 1