Exposé zur Studienarbeit

Phylogenie von Graphen - Ähnlichkeit metabolischer Netzwerke

Christian Brandt

24. Mai 2006
Betreuer: Prof. Ulf Leser

1 Einleitung

In der Biologie werden Interaktionen zwischen den Komponenten des Zellstoffwechsels untersucht. Sie können als Graphen bzw. Netzwerke dargestellt werden. Die Knoten repräsentieren dabei Enzyme, Proteine oder andere Verbindungen, und die Kanten chemische Interaktionen. Traditionell werden einzelne Stoffwechselvorgänge, wie zum Beispiel die Glykolyse oder der Citratzyklus, getrennt voneinander betrachtet. Diese *metabolischen Netzwerke* sind mittlerweile für viele Organismen bekannt.

Nun verändern sich während der Evolution mit der Basensequenz auch die Proteine. Diese zeigen daraufhin ein anderes Bindungsverhalten und damit ändern sich auch die metabolischen Netzwerke. Die Veränderungen in den Netzwerken müssten daher die Abstammungsgeschichte von Arten nachzeichnen.

Um aufgrund der Ähnlichkeit von Netzwerken phylogenetische Bäume zu bauen, ist es notwendig ein Abstandsmaß zwischen ihnen zu definieren. Dafür gibt es verschiedene Möglichkeiten. Man aligniert zum Beispiel Pfade, das heißt einzelne Reaktionsketten innerhalb der Netzwerke [KSK$^+$03] oder auch Bäume [PRYZ05] miteinander und benutzt diese Maße als Approximation des evolutionären Abstandes.

2 Ziel

In der Studienarbeit soll die Hypothese untersucht werden, ob man über Phylogenien von metabolischen Netzwerken tatsächliche Stammbäume rekonstruieren kann. Dazu werden bekannte abstandsbasierte phylogenetische Algorithmen verwendet. Kern der Arbeit wird die Implementation eines Verfahrens zur Bestimmung der nötigen Graphabstände sein. Eine Möglichkeit ist der Graph-Edit-Abstand. Die Berechnung soll exakt durchgeführt werden. Da das Problem eng mit dem NP-harten Graph Transformation Problem

[Lin94] verwandt ist, wird man diese Lösung nur für sehr kleine Netzwerke anwenden können. Die Metrik soll mit anderen Ähnlichkeitsmaßen zwischen Graphen bezüglich der Qualität der damit abgeleiteten Bäume verglichen werden.

3 Vorgehen

Zunächst muss festgelegt werden, wie der Graph-Edit-Abstand genau definiert wird. Im Speziellen geht es darum, welche Transformationsoperationen erlaubt sind und mit welchen Kosten sie verbunden sind. Anschließend muss ein Algorithmus zur exakten Berechnung des Graph-Edit-Abstands ausgearbeitet werden. Folgende alternative Abstandsmaße kommen für den Vergleich in Betracht:

- Hammingabstand der Adjazenzmatrizen der gemeinsamen Knoten und eine noch festzulegende Bestrafung für nicht gemeinsame Knoten
- Die Anzahl nicht gemeinsamer Pfade in beiden Netzwerken

Alle Verfahren werden in Java implementiert. Zur Berechnung und Visualisierung der phylogenetischen Bäume wird die Software SplitsTree verwendet. Als abstandsbasierter Algortihmus bietet sich Neighbor-Joining an. Die Methoden sollen anhand einer manuell ausgewählten Menge Netzwerke aus der Kyoto Encyclopedia of Genes and Genomes [KGH$^+$06] verglichen werden. Bei der Auswahl sind folgende Punkte zu beachten:

1. Welcher Netzwerktyp soll untersucht werden?
   - Protein-Interaktionen in metabolischen Netzwerken
   - chemische Reaktionen in metabolischen Netzwerken (In diesem Fall stellt sich zusätzlich die Frage, wie man die Reaktionen im Graphen repräsentiert.)
   - Protein-Interaktionen in regulatorischen Netzwerken
2. Das Netzwerk darf wegen der Komplexität des Problems nicht zu groß sein. Es werden lediglich einzelne, abgegrenzte „Pathways“ aus KEGG untersucht. Diese enthalten zwischen 20 und 350 Knoten.
3. Das Netzwerk sollte in einer repräsentativen Menge von Organismen vorkommen. Dabei wird das Größenverhältnis zwischen artspezifischem Pathway und KEGG-Referenz-Pathway ausschlaggebend sein.

Die Glykolyse ist ein gutes Beispiel für ein in vielen Organismen konserviertes Netzwerk. Betrachtet man hier die indirekten Protein-Interaktionen und ignoriert Verbindungen zu anderen Netzwerken, so enthält das Referenznetzwerk 40 Knoten. Ein Knoten entspricht einem Enzym. Es gibt eine gerichtete Kante zwischen den Enzymen E_1 und E_2, wenn das Produkt einer Reaktion, die E_1 katalysiert, gleichzeitig Substrat einer Reaktion ist, die E_2 katalysiert. Der resultierende phylogenetische Baum über den ausgewählten Organismen könnte bespielsweise mit dem Baum der NCBI taxonomy oder dem des Tree of Life Projektes verglichen werden.

Literatur

[KGH+06] Minoru Kanehisa, Susumu Goto, Masahiro Hattori et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue):D354–D357, 2006.

[KSK+03] Brian P. Kelley, Roded Sharan, Richard M. Karp et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11394–11399, 2003.

[Lin94] Chih-Long Lin. Hardness of approximating graph transformation problem. In *ISAAC '94: Proceedings of the 5th International Symposium on Algorithms and Computation*, pages 74–82, London, UK, 1994. Springer-Verlag.

[PRYZ05] Ron Y. Pinter, Oleg Rokhlenko, Esti Yeger-Lotem and Michal Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–3408, 2005.