Visuelle Graphanfragen für biologische Netzwerke

Exposé

Stefan Ahl

Betreuer: Prof. Dr. Agnès Voisard
Freie Universität Berlin, Institut für Informatik
Fraunhofer ISST, Mollstr. 1, 10178 Berlin

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

Zeitraum: 23. Januar 2006 bis 23. Juli 2006

Motivation

In der modernen Biologie sind die Signalübertragung innerhalb und zwischen Zellen, die Regulation von Genexpressionen oder die generelle Interaktion zwischen Molekülen wichtige Forschungsgebiete. Deren untersuchte Objekte werden als Netzwerke oder Graphen modelliert. Viele Fragestellungen der Biologie lassen sich auf das Finden von Strukturen oder Mustern innerhalb dieser Graphen zurückführen. Hierfür ist eine einfache aber mächtige Sprache erforderlich, die es dem Anwender unabhängig von seiner speziellen Anwendung erlaubt, nach solchen Strukturen oder Mustern innerhalb eines Graphen zu suchen. Als Beispiel sei hier das in [2] veröffentlichte Teilnetzwerk der Proteininteraktionen von E. coli genannt, das aus über 791 chemischen Verbindungen, organisiert in mehr als 744 biochemischen, enzym-katalysierten Reaktionen besteht.

Am Lehrstuhl für Wissensmanagement in der Bioinformatik der Humboldt-Universität zu Berlin wurde die Sprache PQL, Pathway Query Language, und deren Konzepte entwickelt [3]. PQL findet Teilstrukturen eines Graphen anhand der in einer Anfrage angegebenen Bedingungen. In einer PQL Abfrage werden diejenigen Knoten des Graphen an Knotenvariablen gebunden, die alle Bedingungen der Abfrage erfüllen. Bedingungen an Zusammenhänge zwischen Knoten können aber immer nur für Knotenpaare formuliert werden. Eigenschaften kompletter Pfade kann man damit nicht ausdrücken. PQL ist auf einen Eingabegraphen und die Erzeugung eines Ergebnisgraphen limitiert und bietet noch keine Möglichkeit der Visualisierung. PQL fehlen bisher also noch folgende, wesentliche Fähigkeiten:

- Pfadvariablen als neues Sprachmittel für Eigenschaften kompletter Pfade
- Die parallele Verarbeitung mehrerer Graphen und die Erzeugung mehrerer Ergebnisgraphen
- Die visuelle Spezifizierung einer Anfrage und eine visuelle Ergebnisanzeige

Zielsetzung

Ziel dieser Diplomarbeit ist die theoretische und praktische Erweiterung von PQL um die oben erwähnten, noch fehlenden Sprachfähigkeiten und die Umsetzung der Möglichkeiten zur visuellen Modellierung von PQL Anfragen und der visuellen Anzeige von Anfrageergebnissen. Diese Erweiterungen und deren konkrete Anwendungsmöglichkeiten werden theoretisch erarbeitet, durch praxisorientierte Beispiele ergänzt und sollen zum Abschluss der Diplomarbeit in einer vorführbaren Version implementiert sein.

Vorgehensweise Visualisierung

Für die Visualisierung von Netzwerken wird eine vorhandene Software genutzt. Hierfür eignet sich die Software Cytoscape in besonderem Maße [4]. Es handelt sich um Open Source Software, die Basisfunktionalitäten für das Layout und eine einfache Suche nach Komponenten von Netzwerkgraphen bereitstellt. Der Softwarekern ist durch eine vorhandene Plug-In-Architektur erweiterbar, welche die Integration einer visuellen Modellierung von PQL Statements und deren Überführung in PQL Syntax erheblich erleichtern soll. Die zentrale Organisationseinheit von Cytoscape ist ein Netzwerkgraph, der der relationalen Repräsentation von Graphen für PQL sehr ähnlich ist. Folgende allgemeine Schritte werden künftig bei der visuellen Modellierung und Anzeige in PQL den typischen Arbeitszyklus von der Erstellung einer Abfrage bis zum Anzeigen des Anfrageergebnisses darstellen:

1. Auswahl der Eingabegraphen
2. Visuelle Modellierung eines PQL Statements
3. Übersetzung der visuellen PQL Anfrage in eine entsprechende PQL Syntax
4. Generierung von PL/SQL Prozeduren in Oracle anhand Punkt 3
5. Berechnung und Speicherung von Anfrageergebnissen in der Oracle Datenbank anhand der PL/SQL Prozeduren
6. Visualisierung der Abfrageergebnisse

Vorgehensweise Spracherweiterung

Das PQL zugrundeliegende Modell ist ein Graph. Die augenblickliche Implementierung eines solchen Graphen benutzt ein relationales Schema. Die Trennung von Pfaden und Knoten in eigenen Relationen eignet sich auch für die Einführung von Pfadvariablen. Da die Pfade eines Graphen einmalig vorberechnet werden, sind somit bei der Ausführung von PQL Anfragen alle Pfade des Graphen in der Datenbank für mögliche Bindungen an Pfadvariablen vorhanden.

Die Speicherung von Eingabegraphen, Zwischenergebnissen und Ergebnisgraphen soll analog zur bisherigen Vorgehensweise in Tabellen der Oracle Datenbank erfolgen, um auch die Ausführung einer Abfrage auf mehreren Graphen und die Berechnung mehrerer Ergebnisgraphen zu unterstützen.

Die theoretische Ausarbeitung der Diplomarbeit wird die Erweiterung der PQL Syntax und Semantik um reguläre Pfadausdrücke, ein Modell für visuell erstellte PQL Anfragen und das Mapping zwischen diesen und regulärer PQL Syntax enthalten. Auch für die Überführung von Ergebnisgraphen in die visuelle Darstellung von Cytoscape wird ein Modell erarbeitet. Ein Compiler wird anhand des erarbeiteten Mapping-Modells die Übersetzung einer visuell erzeugten PQL Anfrage in die reguläre Syntax von PQL übernehmen. Das generierte PQL Statement wird dann wie gehabt in eine Oracle PL/SQL Prozedur mit einem größtmöglichen Anteil an SQL übersetzt und dort ausgeführt.

Beispiele für die erweiterte Semantik von Abfragen werden die neuen Möglichkeiten von PQL erläutern.

Im praktischen Teil der Diplomarbeit werden alle obigen Erweiterungen implementiert.

Literatur

[1] Barabasi, A.-L. (2004). "Global organization of metabolic fluxes in the bacterium Escherichia coli", Nature, Vol. 427, 26 February 2004

[2] Karp, Peter D. (2001). "Pathway databases: A Case Study in Computational Symbolic Theories", Science, Vol. 293 no. 5537, pp. 2040 - 2044, 14 September 2001

[3] Leser, U. (2005). "A Query Language for Biological Networks", Department for Computer Science, Humboldt-Universität Berlin,Technischer Report 187.

[4] "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", Genom Research 13:2498-2504