Expóse zur Studienarbeit
"Title still missing"

Betreut durch: Silke Trissl, Prof. Ulf Leser

Philipp Hussels

3. Januar 2007

1 Motivation

Im Bereich der molekularbiologischen Forschung existieren für verschiedene Domänen jeweils mehrere Datenquellen, die sich inhaltlich und strukturell überschneiden, aber in der Methode der Datengewinnung oder der Sicht auf die Daten differieren. Für viele praktische Fragestellungen ist es erforderlich Datenquellen unterschiedlicher Domänen anzufragen und Links zwischen diesen Quellen zu verfolgen. Je nach Auswahl der Quellen aus den einzelnen Domänen können die Anfrageergebnisse variieren. Für einen Benutzer ist dabei nicht ersichtlich, welche der Ergebnisse besonders vertrauenswürdig (konfident) oder signifikant sind.

2 Zielsetzung

Im Rahmen dieser Studienarbeit sollen Methoden zur performanten Berechnung von Werten für Konfidenz und Signifikanz von Anfrageergebnissen entwickelt werden. Diese Methoden sollen durch Implementierung in der integrierten Proteindatenbank Columba[1] evaluiert werden.
Zur Berechnung beider Größen soll die Redundanz der verfügbaren Datenquellen ausgenutzt werden. So ist ein Anfrageergebnis besonders vertrauenswürdig, wenn es durch mehrere Quellen verifiziert wird. Dabei ist zu beachten, dass Datenquellen einer Domäne eine mehr oder weniger große Abhängigkeit aufweisen können. Biologische Daten sind i.A. Ergebnisse experimenteller Arbeit. Ist eine Information beispielsweise in zwei Quellen enthalten, macht es einen Unterschied, ob die Daten dieser Quellen mit der gleichen oder unterschiedlicher experimenteller Methodik erhoben wurden. Vertrauenswürdiger sind Informationen, die von unabhängigen Quellen bestätigt werden. Einen Hinweis auf den Grad der Abhängigkeit zwischen Datenquellen können die Größen von Vereinigungs- und Schnittmengen geben. Unabhängige Datenquellen werden eine gemessen an der Vereinigungsmenge kleine Schnittmenge aufweisen. Auf die Signifikanz einer

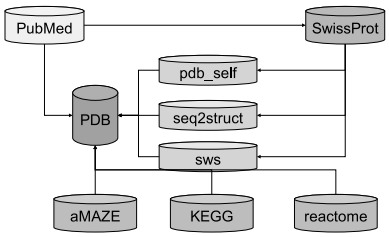

Abbildung 1: Columba Teilschema

Datenquelle für eine gegebene Anfrage kann aus der Zahl ihrer Antworten und der Größe der Datenquelle geschlossen werden. Insbesondere dann, wenn eine kleine Datenquelle viele Ergebnisse liefert, ist davon auszugehen, dass sie signifikante Daten enthält. Ergebnissen, die von solchen Quellen verifiziert werden, muss eine entsprechend hohe Signifikanz zugeordnet werden.

3 Vorhandene Daten

Das für diese Arbeit relevante Teilschema von Columba ist in Abbildung 1 dargestellt. Im Zentrum stehen Strukturen aus der Protein Data Bank[2] (PDB). Um diese herum sind sternförmig Datenquellen aus den Domänen (im Folgenden als Dimensionen bezeichnet) metabolische Pfade, Sequenzen und Publikationen angeordnet. Für die Dimension metabolische Pfade (Pathways) stehen mit KEGG Ligand[3], aMAZE[4] und Reactome[5] drei sehr heterogene Datenquellen zur Verfügung. Um diese Quellen trotz unterschiedlicher Struktur in gleicher Weise anfragen zu können, wurden Informationen jeweils auf Ebene von Reaktionen extrahiert. Edukte, Produkte und Reaktionen sind Knoten eines gerichteten Graphen. Dieser Graph wird in der Datenbank durch eine Kantenrelation abgebildet. Ein Mapping auf PDB-Strukturen ist über die EC-Nummern der Reaktionen und die Link-Quelle PDBSprotEC[6] gegeben.
Als Datenquelle für die Dimension Sequenzen dient Swiss-Prot[7]. Diese Quelle ist über die Link-Quellen Seq2Struct[8], PDBSWS[9] und E-MSD[10] mit der PDB verbunden.
In ähnlicher Weise ist Pubmed als Quelle für Publikationen sowohl über Swiss-Prot als auch direkt mit der PDB verlinkt. Durch den Link über Swiss-Prot besteht eine Abhängigkeit zwischen den Dimensionen Publikationen und Sequenzen, die bei der Berechnung von Konfidenz und Signifikanz berücksichtigt werden muss.
Ein Beispiel für eine Anfrage, deren Ergebnisse schließlich bewertet werden sollen, wäre: “Welche PDB-Strukturen sind sowohl in Swissprot als auch in Pubmed mit dem Keyword ’Glycolysis’ annotiert und weniger als 4 Reaktionsschritte von einem Enzym mir der EC-Nummer ’2.7.1.1’ (Hexokinase) enfernt?”

4 Vorgehensweise

Zunächst werden die Größen Konfidenz und Signifikanz exakt definiert. Auf Basis dieser Definition werden Methoden zur Berechnung beider Größen bei Anfragen gegen eine Dimension entwickelt. Um Performanz zu gewährleisten, werden alle Datenquellen einer Dimension integriert. Zu jeder Dimension wird eine Tabelle erzeugt, die alle Tupel der zugehörigen Datenquellen enthält. Jedem Tupel wird ein Bit-Vector zugeordnet, der beschreibt, in welchen Datenquellen das Tupel enthalten ist. Dadurch entfallen zur Anfragezeit aufwändige Schnittmengenberechnungen.
Können Signifikanz und Konfidenz eines Ergebnisses für Anfragen gegen eine Dimension berechnet werden, soll eine geeignete Aggregatfunktion gefunden werden, um diese Werte für Anfragen gegen mehrere Dimensionen aus den Werten der einzelnen Dimensionen zu berechnen. Schließlich wird ein Set von Anfragen zusammengestellt, um Korrektheit und Performanz der Methoden zu evaluieren.

Literatur

[1] Silke Trissl, Kristian Rother, Heiko Mueller, Thomas Steinke, Ina Koch, Robert Preissner, Cornelius Froemmel, and Ulf Leser. Columba: an integrated database of proteins, structures, and annotations. *BMC Bioinformatics*, 6(1), March 2005.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, January 2000.

[3] S. Goto, Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*, 30(1):402–404, January 2002.

[4] C. Lemer, E. Antezana, F. Couche, F. Fays, X. Santolaria, R. Janky, Y. Deville, J. Richelle, and S. J. Wodak. The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res*, 32(Database issue), January 2004.

[5] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. d. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Supplement 1):D428+, January 2005.

[6] A. C. Martin. PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, 20(6):986–988, April 2004.

[7] E. Gasteiger, E. Jung, and A. Bairoch. SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr Issues Mol Biol*, 3(3):47–55, July 2001.

[8] Allegra Via, Andreas Zanzoni, and Manuela Helmer-Citterich. Seq2Struct: a resource for establishing sequence-structure links. *Bioinformatics*, September 2004.

[9] Andrew C. Martin. Mapping PDB chains to UniProtKB entries. *Bioinformatics*, 21(23):4297–4301, December 2005.

[10] S. Velankar, P. Mcneil, V. Mittard-Runte, A. Suarez, D. Barrell, R. Apweiler, and K. Henrick. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, 33(Supplement 1):D262+, January 2005.