Humboldt-Universität zu Berlin

Wissensmanagement in der Bioinformatik

Vergleich verschiedener Maße für die paarweise Protein-Ähnlichkeit basierend auf Gene Ontology Annotationen

Exposé zur Studienarbeit von Andrej Masula

4. Juni 2008

Betreuer: Philip Groth, Prof. Ulf Leser

1 Hintergrund und Motivation

Gene spielen als Bauplan für die daraus entstehenden Proteine und somit deren Funktion die entscheidende Rolle. In der Genomforschung geht man davon aus, daß Gene mit ähnlicher Sequenz ihrer Nukleotide auch eine ähnliche Funktion repräsentieren. Verfahren, die auf Ähnlichkeitsmaßen basieren sind z.B. das Smith-Waterman- und das BLAST-Verfahren [7][8].
Mittlerweile sind sehr viele Gene auch in ihrer Funktion, genauer gesagt, der Funktion der daraus entstandenen Proteine beschrieben. Das heißt, in Gendatenbanken wie der *Entrez Gene* [6] sind die Gensequenzeinträge durch Annotation erweitert. Dabei wurden die Funktionen beispielsweise experimentell ermittelt.
Im Bereich der Bioinformatik ist das Interesse stark, auf Basis dieser funktionellen Annotationen der Gene die sogenannte *semantische Ähnlichkeit* zu berechnen.

Eine Möglichkeit Wissen strukturiert zu erfassen sind Ontologien. Dazu werden realweltliche Konzepte einer Domäne durch verschiedene Beziehungstypen miteinander verknüpft. Zur Strukturierung des Wissen über Gen- bzw. Protein-Funktionen wurde u.a. die *Gene Ontology (GO)* [2] entwickelt. Genau genommen sind in der *GO* drei Ontologien enthalten, die die Gen-Funktionen vor dem Hintergrund drei verschiedener Aspekte einordnen: *molecular function*, *biological process* und *cellular component*.
In der Anwendung werden Gen-Annotationen mit Konzepten (*Termen*) einer Ontologie assoziiert, so daß das strukturelle Wissen der Ontologie auf das Gen anwendbar ist.

Um semantische Ähnlichkeit zu berechnen, müssen alle, in *GO* assoziierten Terme zweier Gene, miteinander verglichen werden. Dazu werden üblicherweise eine Reihe von Verfahren benutzt, deren Effektivität in dieser Arbeit beurteilt werden soll. Es handelt sich um folgende Verfahren:

- Verfahren der mittleren Ähnlichkeit nach Francisco M. Couto et al. [1]
- Verfahren der mittleren Ähnlichkeit nach P. Lord et al. (2003) und Haiying Wang et al. (2004) [5],[4].
- Ermittlung der Ähnlichkeit mittels Bildung reziproker Term-Paare nach Ying Tao et al. (2007) [9].
- Maximum-Ähnlichkeitsmaß nach Xiang Gou et al. (2006) [3].

Alle genannten Verfahren bauen auf der Berechnung der semantischen Ähnlichkeit zweier einzelner Terme auf. Dazu werden in der Literatur oft die drei Verfahren von *Resnik (1999)*, *Jiang und Conrath (1997)* und *Lin (1998)* erwähnt.
Diese Verfahren nutzen jedoch nicht nur die strukturellen Informationen der Ontologie, sondern kombinieren zusätzlich mit einem statistischen Maß, dem Informationsgehalt. Der Informationsgehalt eines Terms ist umso höher, je geringer seine Auftretenswahrscheinlichkeit in einem zugrundeliegenden Korpus (z.B. Gendatenbank) ist.

2 Ziele

Ziel dieser Arbeit ist, die oben genannten Verfahren zur Berechnung der semantischen Ähnlichkeit von Genen, respektive Proteinen zu vergleichen. Als Ergebnis dieses Vergleichs soll ein Benchmark entstehen, in dem die Verfahren bezüglich ihrer Effektivität eingeordnet sind. Die Effektivität der Verfahren wird daran gemessen, wie gut deren Ergebnisse mit der Sequenz-Ähnlichkeit nach Smith-Waterman korrellieren.
Datengrundlage ist eine Menge von Gensequenzdaten aus den *Entrez Gene Database* und deren Verweise auf Terme der Gene Ontology, sowie die *Biological Process-Ontologie* des Gene Ontology Projekts.

3 Vorgehensweise

Um die Berechnung der Ähnlichkeiten durchzuführen wird eine Software erstellt, in der die Algorithmen der betrachteten Verfahren implementiert sind. Dazu muß die Software die betrachtete Ontologie einlesen und im Speicher bereitstellen können.

3.1 Reduktion zu einer geeigneten Datenbasis

Die Betrachtung aller Gensequenzen aus der Entrez-Gendatenbank ist aufgrund von Kapazitätsbeschränkungen der verfügbaren Hardware nicht möglich. Zur Analyse werden nur die Terme der GO-Ontologie *Biological Process* und nur Gensequenzen mit mindestens 10 Verweise auf GO-Terme herangezogen.
Die mit GO-Termen annotierten Gen-ID's sind in einer flach strukturierten Textdatei bereitgestellt. Daraus werden nur die Gen-ID's selektiert, die obige Bedingungen erfüllen. Die dazugehörigen Gensequenzen werden aus der Entrez-Gendatenbank abgefragt und in einer Textdatei abgelegt.

3.2 Verarbeiten der Ontologie

Die *Biological Process*-Ontologie wird eingelesen und ein entsprechender Graph im Speicher aufgebaut. Die Ontologien des Gene Ontology-Projekts entsprechen in ihrer Datenstruktur einem gerichtetem azyklischem Graphen (DAG: Directed Acyclic Graph). Das bedeutet unter anderem, daß ein Knoten auch mehrere Elternknoten haben kann.
Um die spätere Rechenzeit zu optimieren, wird in diesem Schritt zusätzlich zur Elternknoten-Kindknoten-Beziehung auch eine Vorfahr-Nachfahr-Beziehung gespeichert, da diese in den Algorithmen häufig gebraucht wird.

3.3 Vorverarbeiten der Gendaten

Damit die semantischen Ähnlichkeiten berechnet werden können, muß der Informationsgehalt jedes Terms des Ontologiegraphen ermittelt und dem Term zugeordnet werden. Dazu werden die mengenmäßig reduzierten und annotierten Gendaten von der Software eingelesen und die Auftretenswahrscheinlichkeiten der Terme bestimmt. Weiterhin werden für jedes Gen alle GO-Term-Verknüpfungen in einer Datenstruktur gespeichert.

3.4 Berechnen der Ähnlichkeiten für alle Genpaare

Zu jedem Genpaar wird sowohl die Sequenz-Ähnlichkeit nach Smith-Waterman als Referenzwert ermittelt, als auch die semantischen Ähnlichkeiten gemäß der oben genannten Verfahren. Bei allen semantischen Verfahren kommt zur Berechnung der Term-Term-Ähnlichkeit einheitlich das Verfahren von *Lin* zum Einsatz. Alle Ähnlichkeitswerte werden in einer geeigneten Datenstruktur zum entsprechenden Entrez Gen-ID-Paar gespeichert und in einer Text-Datei zeilenweise ausgegeben.

3.5 Auswertung der Ergebnisse

Zur übersichtlicheren graphischen Darstellung werden die Ergebnistupel der betrachteten Genpaare entsprechend ihrer Sequenz-Ähnlichkeit absteigend geordnet. Gemäß dieser Sortierung kommen zusätzlich alle semantischen Ähnlichkeitswerte in einem Diagramm zur Ausgabe.
Es werden zusätzlich die Korrelationskoeffzienten zwischen jedem semantisch Ähnlichkeits-Algorithmus und dem syntaktischen Referenzverfahren berechnet.

Literaturverzeichnis

[1] Francisco M. Couto, Mario J. Silva, and Pedro M. Coutinho. Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, 61(1):137–152, April 2007.

[2] GO Consortium. *The Gene Ontology (GO).* `http://www.geneontology.org/GO.doc.shtml`.

[3] Xiang Guo, Rongxiang Liu, Craig D. Shriver, Hai Hu, and Michael N. Liebman. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22(8):967–973, 2006.

[4] Oliver Bodenreider Joaquin Dopazo Haiying Wang, Francisco Azuja. Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships. *Proceedings of CIBCB 2004*, 2004:25–31, 2004.

[5] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[6] National Center for Biotechnology Information. *Entrez Gene Database*, 2007. `www.ncbi.nlm.nih.gov/sites/entrez?db=gene`.

[7] Waterman MS Smith TF. Identication of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981. `http://gel.ym.edu.tw/~chc/AB_papers/03.pdf`.

[8] David Lipman Stephen Altschul, Warren Gish. *BLAST.* National Center for Biotechnology Information, 1990. `www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html`.

[9] Ying Tao, Lee Sam, Jianrong Li, Carol Friedman, and Yves A. Lussier. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, 23(13):i529–538, 2007.