Fachgebiet Wissensmanagement in der Bioinformatik

Aufspürung von Gene Ontology Termen in wissenschaftlichen Artikeln

Exposé zur Diplomarbeit

Nikolay Damyanliev
damyanli@informatik.hu-berlin.de
Matr.-Nr. 173593

Betreuer: Prof. Ulf Leser

29. April 2009

1 Hintergrund und bisherige Forschung

Mit der immer wachsenden Menge an Information über Proteine und Gene in molekularbiologischen Artikeln wird die Entwicklung von effektiven Informationsextraktionsverfahren immer wichtiger. Die Information, die sich auf die Funktion der Proteine bezieht, wird meistens manuell durch Kuratoren aus verschiedenen wissenschaftlichen Literaturquellen extrahiert und in biologischen Datenbanken gespeichert [1], um danach eine wichtige Rolle bei vielen Analyseverfahren in der Bioinformatik zu spielen. Gene Ontology (GO) ist das am meisten verbreitete kontrollierte Vokabular zur konzeptionellen Beschreibung von Funktionen von Genprodukten und deren Annotation in Datenbanken [2]. Obwohl die manuelle Annotation sehr genau ist [3], ist sie auch ziemlich zeitraubend und langsam. Dazu überschreitet auch die Wachstumsrate der Anzahl biologischer Artikel die Möglichkeiten für manuelle Annotation. Deswegen ist die Entwicklung von guten automatisierten Annotationswerkzeugen für Gene mit GO-Termen ein aktuelles Thema.

Ein eng mit der Annotation verbundenes Problem ist die Suche nach konkreten Nachweisstellen in den Artikeln für die Annotation der Gene mit GO-Termen. Die Suche nach konkreten Textstellen, die Nachweise für eine Annotation enthalten können, kann die manuelle Annotation oder die Prüfung der automatischen Annotation deutlich erleichtern, indem die Kuratoren, die die Annotation prüfen, nicht die ganzen Artikel durchlesen müssen, sondern nur die vorgeschlagenen Nachweisstellen.

Mit diesen beiden Problemen haben sich mehrere Bioinformatik-Gruppen bei dem Wettbewerb BioCreAtIvE I beschäftigt. Beide Probleme waren als Teilaufgabe 2.2 bzw. Teilaufgabe 2.1 zur Lösung gestellt. Die Gruppen haben zur Lösung mit unterschiedlichem Erfolg verschiedene Text-Mining-Methoden ausprobiert. Ray et al. [6] konstruieren z.B. für jeden GO-Term eine Menge von „Informationstermen“, wobei sie die Abstracts von Artikeln benutzen, die in verschiedenen Datenbanken mit GO-Annotationen als relevant für den jeweiligen GO-Term angegeben wurden. Diese Informationsterme werden dann verwendet, um die Abschnitte mit der höchsten Relevanz zu bestimmten GO-Termen in den Artikeln zu finden, die in BioCreAtIvE als Ausgangsdaten vorgegeben waren. Vespoor et al. [4] konstruieren mithilfe der bereitgestellten Trainingsdaten zu jedem GO-Term Wortmengen, die dann nach TFIDF mit einer Gewichtung versehen werden. Danach gehen sie die Paragraphen in den Testartikeln durch und bilden den Schnitt zwischen den Wörtern im Paragraphen und den konstruierten Wortmengen. Dieser Schnitt gilt danach als Relevanzfaktor zwischen dem Paragraphen und dem GO-Term. Krallinger et al. [5] betrachten nicht die einzelnen Paragraphen als Nachweistexte, sondern immer 3-4 aufeinander folgende Sätze im Text. Dies ermöglicht es, als Resultat konkretere Nachweisstellen zu bekommen – mit der Idee, dass die Funktion eines Genproduktes meistens in 3-4 zusammenhängenden Sätzen beschrieben wird. So werden genauere Stellen und nicht ganze Paragraphen als Resultate ausgegeben.

Im Allgemeinen können diese Methoden in zwei Gruppen geteilt werden – „pattern matching“ und „machine learning“ Ansätze. Beide benutzen die Häufigkeit der Wörter im GO-Vokabular als Grundstein ihrer Bewertungsfunktionen.

2 Zielsetzung

Diese Diplomarbeit baut auf einer Studienarbeit zum Thema „Suche nach Nachweisstellen in Artikeln für vorhandene GO-Annotationen“ auf. In diesem Zusammenhang verfolgt sie die folgenden Hauptziele:

- Verbesserung der Suche nach Nachweisstellen in Artikeln für GO-Annotationen mit verschiedenen Ansätzen (genaue Fehleranalyse der Ergebnisse der Studienarbeit, Einführung von Stemming u.a.)
- Erweiterung der Aufgabe der Studienarbeit auf allgemeine GO-Term-Aufspürung in Artikeln
- Entwicklung eines webbasierten Tools zur Suche nach GO-Termen in PubMed-Abstracts bzw. in ganzen Artikeln. Der eigentliche Algorithmus wird auch als stand-alone Web Service verfügbar sein

Verschiedene Werkzeuge werden ausprobiert, um ihr Nutzen für die Lösung der beiden Probleme zu bestimmen – NLP-Tools zum Thema Konzeptfindung, Stemming von Wörtern, Sentence-Splitter, Benutzung von Wortphrasen mit Beachten der Reihenfolge der Wörter vs. „bags of words“ u.a.

3 Herangehensweise

3.1 Ansätze

Wie in der Studienarbeit, auf der diese Diplomarbeit aufbaut, wird für die generelle Aufspürung der GO-Terme im Artikeltext mithilfe der GO-Datenbank (http://www.geneontology.org/) und weiteren Quellen wie PubMed-Abstracts für jeden GO-Term eine Menge von gewichteten relevanten Wörtern gebildet („bag of words“ oder GO-Wolke). Die Gewichtung der Wörter hängt von deren Häufigkeit sowohl in der GO-Wolke des GO-Terms, als auch in der Menge aller GO-Wolken ab. Die Artikel mit den vermuteten Quellen für GO-Annotation werden dann sukzessiv nach den Wörtern aus den GO-Wolken durchgesucht und danach werden die relevantesten GO-Terme mit deren „Fundstellen“ zurückgeliefert. Bei der Suche nach den Wörtern aus den GO-Wolken wird auf Sentence-Splitter, Stemming und weitere NLP-Tools zurückgegriffen. Die generelle Vorgehensweise wird die Bildung verschiedener GO-Wolken mit Ausprobierung von Einfügen relevanter Wörter aus verwandten GO-Termen und relevanten PubMed-

Abstracts, um deren Nützlichkeit zu bestimmen. Die Nützlichkeit von Stemming wird auch geprüft.

Eine Verbesserung der Ergebnisse der Studienarbeit wird bei der Benutzung von Stemming und spezielleren Word-Splitter (die für biologische Texte angepasst wurden) erwartet. Außerdem wird in dieser Diplomarbeit ein generelles Wörterbuch mit Wörtern und Wortphrasen aus den GO-Wolken aller GO-Terme gebaut (d.h. ~27 000 Terme anstatt 1500 wie in der Studienarbeit). Es wird auch untersucht, ob eine kürzere Distanz zwischen zwei Elementen mit großem Gewicht aus einer GO-Wolke eine bessere Fundstelle für den GO-Term bedeutet.

Für die Suche der Nachweisstellen in Artikeln zu einem bestimmten Protein und dessen Annotation werden auch Tools für Suche von Proteinnamen im Text benutzt (wie z.B. BANNER [7]).

Als Nachweis- bzw. Fundstelle wird immer ein Fenster (in variabler Größe) von aufeinander folgenden Sätzen betrachtet.

3.2 Daten

Ausgangsdaten sind Daten aus der GO-Datenbank (Daten mit Namen, Ontologien von GO-Termen mit Verweisen auf relevante PubMed-Artikel), der UniProt-Datenbank (Daten mit Proteinnamen), der PubMed-Datenbank (Abstracts von Artikeln) und die bei BioCreAtIvE I im Task 2.1 und Task 2.2 benutzten Daten. Die Daten aus BioCreAtIvE I bestehen aus Volltextartikeln, sowie GO-Termen und Proteinen, die in diesen Artikeln nachgewiesen wurden. Die Daten stehen als XML- bzw. HTML-Dateien bereit, d.h. eine Präprozessingphase für die Entfernung aller unnötigen Tags u.a. ist erforderlich.

3.3 Evaluation

Die Ergebnisse dieser Arbeit werden anhand der von Kuratoren bewerteten Resultate von BioCreAtIvE I evaluiert, da diese die einzigen verfügbaren Daten sind, die manuell geprüft wurden. Bei den in dieser Arbeit gelieferten Nachweisstellen kann sich die Formatierung des Textes und die Auswahl bzw. die Größe der Ausschnitte von den Ergebnissen von BioCreAtIvE I unterscheiden – das kann die Evaluation beeinflussen. Ein weiteres Problem ist es, dass nicht alle Resultate bei BioCreAtIvE I bewertet wurden, d.h. für die Resultate, für die es keinen Vergleichstext aus BioCreAtIvE I gibt, kann keine Aussage für deren GO-Relevanz gemacht werden.

Für die Evaluation der allgemeinen GO-Term-Suche gibt es mehr Quellen – die Zusammenhänge zwischen GO-Term und Artikel aus den BioCreAtIvE Daten können ausgenutzt werden, und die Ausgangsdaten von Cakmak et. al. [1] könnten auch von Nutzen sein.

3.4 Laufzeit

Besonders die generelle Suche nach GO-Annotationen im Text kann viel Laufzeit beanspruchen, da nach allen GO-Termen durchgesucht wird. Es ist also angebracht, diese Suche zu optimieren. Dazu können Keywordtrees benutzt werden, mit denen nach mehreren Tokens simultan gesucht werden kann. Außerdem wird bei mehreren GO-Termen immer nach den gleichen Wörtern gesucht – da kann ein großes allgemeines Wörterbuch von Hilfe sein. Im allgemeinen, da es immer nach Strings gesucht wird, werden vermutlich alphabetisch geordnete Baumstrukturen für das Durchsuchen angebracht.

Literatur- und Quellenverzeichnis

[1] A. Cakmak, G. Ozsoyoglu: Annotating Genes Using Textual Patterns, Pacific Symposium on Biocomputing 12, pp. 221-232, 2007

[2] The Gene Ontology Consortium: The Gene Ontology (GO) Database and Informatics Resource, Nucleic Acids Research 32, D258-D261, 2004

[3] V. Lee, E. Camon, E. Dimmer, D. Barrell, and R. Apweiler: Who tangos with GOA? Use of Gene Ontology Annotation (GOA) for biological interpretation of '-omics' data and for validation of automatic annotation tools, Silico Biology, vol. 5, no. 1, pp. 5–8, 2005

[4] Verspoor K, Cohn J, Joslyn C, Mniszewski S, Rechtsteiner A, Rocha LM, Simas T: Protein annotation as term categorization in the gene ontology using word proximity networks; BMC Bioinformatics. 2005;6 Suppl 1:S20

[5] Krallinger M, Padron M, Valencia A: A sentence sliding window approach to extract protein annotations from biomedical articles; BMC Bioinformatics. 2005;6 Suppl 1:S19

[6] Ray S, Craven M: Learning statistical models for annotating proteins with function information using biomedical text; BMC Bioinformatics. 2005;6 Suppl 1:S18

[7] BANNER Named Entity Recognition System - http://banner.sourceforge.net/