Verknüpfung funktioneller Annotationen von Genprodukten mit Nachweistexten in relevanten wissenschaftlichen Artikeln

Nikolay Damyanliev, Humboldt Universität zu Berlin, 20. Juni 2008

Betreuer: Prof. Ulf Leser

Planiertes Abgabedatum: 31. Oktober 2008

Problemstellung

Die Annotation von Proteinen mit GO-Termen (Gene Ontology) ist eine komplexe Aufgabe und es wird viel daran gearbeitet und geforscht. Da die manuelle Annotation sehr zeitraubend ist, werden Proteine oftmals auch nicht überwacht mit automatisch generierten Annotationen versehen. Diese sind aber in vielen Fällen mehr oder weniger ungenau, weswegen eine Untersuchung und Nachprüfung mit Literaturquellen sinnvoll ist. Außerdem wird bei manueller Annotation auf ganze Artikel als Nachweisquellen verwiesen, wobei es von Interesse ist die genauen Nachweisstellen in den Artikeln selbst zu wissen. Die Lösung dieses Problems kann bei guten Resultaten auch als Basis für eine bessere automatisierte GO-Annotation benutzt werden.

Die genaue Problemstellung kann kurz auf die folgende Weise zusammengefasst werden:
Gegeben ein Protein, eine GO-Annotation und ein relevanter wissenschaftlicher Artikel soll der Nachweistext in dem Artikel gefunden werden, der das Protein mit der GO-Annotation in Verbindung bringt.

Hauptprobleme

Das Hauptproblem dieser Aufgabe besteht im Finden des Textes, der die GO-Annotation des Proteins beschreibt. Während das Protein selber, unter Angabe seines Namens und mit Benutzung von Proteinwörterbücher mit Synonymen (egal ob fertigen oder selber aus Proteindatenbanken erstellten) und da die Suche in relevanten Artikeln stattfindet, nicht so schwierig im Text nachzuweisen sollte, sind die GO-Terme mehr eine konzeptionelle Funktionsbeschreibung des Proteins und nicht für Benutzung in wissenschaftlichen Texten gedacht. D.h. diese stehen sehr oft nur als Beschreibung eines Konzepts da und/oder mit anderen Worten beschrieben, was deren Nachweis sehr schwierig macht. Das bedeutet, die Aufgabe, worauf man sich konzentrieren muss, besteht darin, die GO-Annotation an den fließenden Text näher zu bringen (oder auch umgekehrt). Da dieses aber sehr viel mit Semantik der Wörter verbunden ist, muss dieser Zusammenhang künstlich „konstruiert werden“. Ideen dazu sind die Normalisation des Textes (Annäherung des Textes an den GO-Term) oder auch Suche nicht nur nach GO-Termen im Text, sondern auch nach mit einem GO-Konzept verbundenen weiteren Begriffen, die man z.B. aus Ontologien oder externen Datenbanken extrahieren kann (Annäherung des GO-Terms an den Text durch Bildung einer „GO-Wolke“ mit relevanten Phrasen und Begriffen).

Bisherige Forschung

Diese Aufgabe wurde beim BioCreAtIvE I Wettstreit (2004) als Teilaufgabe 2.1. von ca. 20 Teilnehmergruppen bearbeitet. Nach den nicht so guten Resultaten bei dem Wettbewerb (viele Teilnehmer berichten über mehr oder weniger enttäuschende Werte Precision/Recall) wird in dem Gebiet auch weitergeforscht. Der GOAnnotator von [1] macht einen guten Eindruck mit einer Precision von 93% (leider aber auf zu wenigen Daten getestet). Es wird

die Ontologiestruktur als Basis benutzt, wobei durch die Beschreibung der GO-Terme die Ähnlichkeit zwischen diesen und auch deren einzelne Auftretenswahrscheinlichkeit berechnet wird, und damit wird auch entschieden an welchen Stellen welche Terme gemeint werden könnten. Aus [2] gewinnt man auch interessante Information – es werden auf verschiedenen GO-Termen trainierte SVMs benutzt, um dann die unterschiedlichen Textbereiche nach Relevanz zu klassifizieren. Die Resultate sind nicht so beeindruckend wie die Beobachtung, dass eine SVM, die auf einem generelleren GO-Term trainiert und auf einen spezielleren GO-Term angewendet wurde, nicht viel schlechtere Resultate erzielt als so eine, die gleich auf dem spezielleren trainiert wurde.
Als Nachweistext werden von den verschiedenen Gruppen auch verschiedene Textbereiche benutzt – am häufigsten werden Absatze genommen, ebenso können aber auch einzelne Sätze, ein Paar oder eine größere Menge von nacheinander folgenden Sätzen benutzt werden.

Daten und Evaluation
Die Daten bestehen aus drei Quellen.

1. Relevante Artikel, in denen die Verbindung zwischen dem Protein und dem GO-Term gesucht wird – werden für Forschungszwecke auf der BMC-Webseite in Fulltext in XML-Format bereitgestellt. Es werden nur diese Artikel betrachtet, auf denen die Teilnehmer an BioCreative I Task 2.1. gearbeitet haben. Diese beinhalten neben dem Text auch lauter XML-Tags, die meistens keine Relevanz zum Text haben und rausgefiltert werden sollen.

2. GO-Ontologien sind auf der Seite von EBI über das GOA-Projekt als Text-Dateien erhältlich. Diese enthalten für jeden GO-Term die Ontologie, in der er auftaucht (molecular function, biological process, cellular component), die Swiss-Prot-AN und Swiss-Prot-ID des Proteins, mit dem er verbunden ist und die PubMed-ID des Artikels, auf dessen Basis die Verbindung registriert wurde.

3. Testdaten werden als XML-Datei, die die durch Kuratoren manuell geprüften und evaluierten Resultate der Teilnehmer an dem BioCreative I Task 2.1 enthält. Das ist die gleiche Information wie in der GO-Ontologie-Datei, aber bei jedem Eintrag steht auch die Stelle aus dem PubMed Artikel dabei, die die Verbindung des Proteins und des GO-Terms nachweist, und auch die Bewertung des Kurators zur Relevanz dieser Stelle.

Da die Evaluation nur auf kurierten Daten möglich ist, sind Eingangsdaten genau die Tripel (Protein, GO-Term, PubMed-ID eines relevanten Artikels), die aus den Testdaten aus 3. extrahiert werden. Mit der Hilfe der dazugehörigen Synonymliste des Proteins und der GO-Wolke wird der Artikel analysiert und es wird eine Textstelle (ca. 4 Sätze) ausgegeben. Ausgabe ist dann eine XML-Datei, ähnlich zu der Datei mit den Testdaten – das Tripel mit den Eingangsdaten wird um die oben beschriebene Textstelle erweitert. Die Ausgabedatei wird dann mit der Testdatei aus 3. verglichen und mit Hilfe der Bewertungen der Kuratoren evaluiert.
Zur Bildung von Synonym- und Abbriviaturlisten für einzelne Proteine wird die Swiss-Prot Datenbank benutzt.
Zur Bildung einer GO-Wolke wird die GO-Ontologie benutzt, in der der GO-Term ist, und zwar werden allgemeinere GO-Terme (bis zu einer bestimmten Distanz) in die Wolke aufgenommen und speziellere nicht (dazu werden Beziehungen wie „is_a“ oder „relationship“ in der Ontologieninfo über das GO betrachtet). Weiter können auch Flexione oder Verbalisierungen mancher Wörter hinzugefügt werden. Andere relevante Erweiterungen der

GO-Wolke, z.B. durch die Ausnutzung der Zusammenhänge zwischen GO-IDs, PubMed-IDs und Swiss-Prot-IDs (d.h. zwischen GO-Termen und PubMed-Abstracts), werden auch berücksichtigt. Alle Wörter in der GO-Wolke werden nach Wichtigkeit, Relevanz und Informationsgehalt gewichtet.

Lösungsansätze

Wie schon beschrieben werden zuerst zu dem gegebenen Protein die Synonym- und Abbriviaturliste und zu dem gegebenen GO-Term die GO-Wolke erstellt. Außerdem werden die Artikel nur auf Haupttext gekürzt und fast alle XML-Tags werden verworfen (bis auf die Paragraph-Tags und andere, die bei der Suche ausgenutzt werden können). Dieser Prozess zählt zur Aufbereitung der Daten und wird bei der Laufzeit nicht mitgezählt.

Für die Suche nach relevanten Nachweistextstellen müssen zwei Probleme gelöst werden – (1) Suche nach Erwähnungen vom Protein im Text, (2) Suche nach Erwähnungen vom GO-Term. Die Relevanz kann für beide Suchterme separat berechnet werden und dann geeignet aggregiert werden, z.B. durch eine Formel (z.B. $Relevanz_{Protein}$ * $Relevanz_{GO\text{-}Term}$), die die beste Textstelle für den Nachweis des Zusammenhangs zwischen den beiden berechnet. Da das Finden des GO-Termes im Text schwieriger ist, kann seiner Relevanz ein größeres Gewicht gegeben werden. Eine Mindestgrenze für jede der Relevanzen kann auch gestellt werden.

Da für diesen Zusammenhang zwischen Protein und GO-Term eine genaue Stelle im Text gesucht wird, sollten die beiden im Text nicht fern (im Sinne von Wort- und Satzdistanz) voneinander auftreten. Da ein Absatz einen Sinnzusammenhang bezeichnet und mit dem Ende des Absatzes ein bestimmtes Gedanke oder Thema endet, wird vermutet, dass der Zusammenhangsnachweis in einem Paragraph (und zwar in nacheinander folgenden Sätzen) gesucht werden sollte. Da einige Paragraphe in den Literaturquellen ziemlich lang sind, ist es sinnvoll nur einen Teil des Paragraphs als Nachweistext zurückzugeben. Ein guter Ansatz wäre also vielleicht ein Sentence-Sliding-Window von 4-5 Sätzen durch den größeren Paragraphen zu schieben (wie in [3]). Die Sätze können z.B. mit dem OpenNLP SetnenceDetector (http://opennlp.sourceforge.net/api/index.html) identifiziert werden.

Die Relevanzmessung der Sätze kann z.B. nach Smith-Waterman-Distanz zwischen den Wörtern des Satzes und diesen aus der Proteinsynonymliste bzw. der GO-Wolke berechnet werden. Vorher werden auch alle Wörter im Satz nach Informationsgehalt durch TF/IDF gewichtet. Dann wird aus der Relevanz der einzelnen Sätze die Relevanz des Textbereiches gebildet (bei zu kleinem Relevanzgewinn beim Einfügen des letzten/ersten Satzes in den Bereich kann dieser auch nicht betrachtet werden). Zum Laufzeitgewinn kann der Sliding-Window speziell für Proteinsuche auf 1 Satz verkleinert werden – da der Proteinname eigentlich ganz (oder als Abbriviatur) in einem Satz vorkommen sollte, wird kein Relevanzverlust erwartet.

Literaturquellen

[1] Finding genomic ontology terms in text using evidence content Francisco M Couto, Mário J Silva, Pedro M Coutinho BMC Bioinformatics 2005, 6(Suppl 1):S21 (24 May 2005)

[2] Mining protein function from text using term-based support vector machines Simon B Rice, Goran Nenadic, Benjamin J Stapley BMC Bioinformatics 2005, 6(Suppl 1):S22 (24 May 2005)

[3] A sentence sliding window approach to extract protein annotations from biomedical articles Martin Krallinger, Maria Padron, Alfonso Valencia BMC Bioinformatics 2005, 6(Suppl 1):S19 (24 May 2005)