Exposé zur Diplomarbeit

Rekonstruktion biologischer Pathways aus sehr großen Korpora

Franziska Brosy

18. November 2009

Betreuer: Philippe Thomas, Prof. Dr. Ulf Leser

Hintergrund

In den Lebenswissenschaften ist das Erforschen des Verhaltens von Zellen, Gewebe oder Organismen, in Hinblick auf ihre wesentlichen molekularen Bestandteile eine der wichtigsten Herausforderungen [Junker et al., 2008]. Im Besonderen bildet das Verständnis der biologischen Prozesse eine Grundlage für maßgeschneiderte Diagnostik mit Auswirkungen auf mögliche Therapien für Krankheiten wie beispielsweise Krebs [ColoNET, 2009].

Eine Vielzahl biologischer Prozesse sind in das abstrakte Konzept eines komplexen Netzwerkes übersetzbar. Dadurch werden biologische Probleme mathematisch nachvollziehbar und somit eine Analyse der Netzwerkeigenschaften, für ein besseres Verständnis des Aufbaus zellulärer Funktionen, möglich [Junker et al., 2008]. Einen funktionsbezogenen kleineren Ausschnitt eines Netzwerkes nennt man Pathway. Genauer formuliert sind Pathways eine Folge biochemischer Reaktionen und repräsentieren Beziehungen zwischen biomedizinischen Entitäten wie Gene oder Proteine. Unter all diesen Beziehungen sind die Interaktionen zwischen Proteinen die grundlegendsten für viele biologische Prozesse [Yao et al., 2004]. Diese Interaktionen können in Form eines komplexen Netzwerkes dargestellt werden, indem die Proteine die Knoten und die Interaktion zwischen zwei Proteinen die Kanten repräsentieren [Özgür et al., 2008].

Das Wissens über Protein-Protein-Interaktionen (PPI) wird in den Lebenswissenschaften, wie vieles weitere biomedizinische Wissen auch, zu einem großen Anteil in Textform aufgezeichnet.

Das heißt, experimentelle Ergebnisse werden in wissenschaftlichen Artikeln veröffentlicht. Diese Ergebnisse sind bspw. über den Online-Literaturservice PubMed[1] zugänglich. Die Anzahl der in PubMed veröffentlichten Beiträge steigt exponentiell. Aus der Literatur manuell extrahiertes Wissen über u.a. PPIs wird in öffentlichen Datenbanken gespeichert. Die manuell extrahierten Informationen decken jedoch nur einen kleinen Teil der Informationen ab, die in der schnell wachsenden Menge der biomedizinischen Literatur zu finden sind [Özgür et al., 2008].

Zur automatischen Extraktion von PPIs aus biomedizinischen Texten existieren bereits viele Tools. Die PPI-Extraktionsmethoden, oft auf Satzebene angewandt, erzielen eine gute Precision (bis zu 95%) aber einen geringen Recall [Skusa et al., 2005]. Die Kennwerte Precision und Recall wurden zur Evaluierung der Methoden auf Gold-Standard Korpora[2] berechnet. [Giles et al., 2008] wendeten ihr Verfahren zur Relationsextraktion auf einem Gold-Standard Korpus (GENIA, ca. 2000 Abstracts) und einem großen nicht Gold-Standard Korpus (verfügbare MEDLINE Abstracts - ca. 18 Mio.) an. Für den Gold-Standard Korpus wurden hohe Precision- und Recallwerte ermittelt, aber das Ergebnis auf dem großen Korpus führte zu einer hohen Falsch-Positiv-Rate.

Des Weiteren gibt es bereits Tools, die basierend auf den extrahierten PPIs, die Beziehungen in Netzwerken zusammenfassen und diese dann graphisch präsentieren [Pavlopoulos et al., 2008]. Einige wenige Ansätze beschäftigen sich mit der ergebnisorientierten (nicht biomedizinisch inhaltlichen) Analyse der Netzwerke. [Chen et al., 2004] entwickelten einen NLP[3]-basierten Ansatz namens Chilibot zur Extraktion biologischer Interaktionsnetzwerke aus Texten. Sie untersuchten 4.584 Abstracts von PubMed und stellten fest, dass die Verbindungen im Netzwerk einem Power Law (viele Knoten mit wenigen Verbindungen und umgekehrt) folgen. [Özgür et al., 2008] erstellten ein krankheitsspezifisches Geninteraktionsnetzwerk ausgehend von 48.245 Artikeln von PubMed Central und ermittelten ein Ranking für die interagierenden Gene basierend auf Verwandtschaftsgrad, Eigenvektoren und Zentralitätsmaßen.

Laut [Skusa et al., 2005] sind die extrahierten Netzwerke aus biomedizinischer Literatur meist unvollständig und können Falsch-Positive enthalten. Auch [Giles et al., 2008] berichten von einer zunehmenden Menge widersprüchlicher Interaktionen und Falsch-Positiven in den Extraktions-

[1] PubMed umfasst derzeit mehr als 19 Millionen Verweise auf biomedizinische Artikel (Stand Oktober 2009; Quelle http://www.ncbi.nlm.nih.gov/pubmed/).

[2] Ein Korpus ist hierbei eine Ansammlung biomedizinischer Texte. Gold-Standard bedeutet, dass die Texte bzgl. bestimmter Textinhalte (bspw. Proteine und Therme, die eine Interaktion zwischen Proteinen beschreiben) manuell annotiert oder automatisch annotiert und manuell nachkorrigiert wurden.

[3] Natural Language Processing.

ergebnissen. Dies deckt sich mit der Aussage von [Yuryev et al., 2006], dass viele Text Mining Verfahren zur Extraktion biologischer Beziehungen sowohl eine hohe Falsch-Positiv-Rate als auch Redundanz in den extrahierten Entitäten hervorbringen.

Viele Ansätze haben bisher ihren Fokus auf relativ kleine Korpora gerichtet. Folglich fehlt eine Überprüfung dieser Ansätze hinsichtlich ihrer Performance auf großen Datensätzen wie bspw. einige Millionen Texte [Giles et al., 2008]. Ferner gilt es aus den extrahierten Netzwerken die relevanten und zuverlässigen Pathways zu filtern und die Ergebnisse auf ihre Übereinstimmung mit bereits bekanntem Wissen zu überprüfen.

Ziele

Ziel dieser Arbeit ist das Rekonstruieren zuverlässiger PPI-Pathways. Ausgehend von einem großen Korpus biomedizinischer Literatur (mehrere Millionen Texte) sollen, unter Verwendung bekannter Methoden, PPIs extrahiert werden. Die extrahierten PPIs sollen dann in einem Netzwerk zusammengefasst werden. Aus diesem Netzwerk wiederum sollen die relevanten PPI-Pathways herausgefiltert werden. Für den Filtervorgang sollen geeignete Rankingmethoden für die Proteine und die PPIs im Netzwerk entwickelt werden. Das primäre Ziel dabei ist, möglichst viele korrekte Pathways unter den gefilterten Pathways zu erhalten. Ein relativ hoher Recall der Pathways müsste sich über die Vielzahl der Textdaten automatisch ergeben. Es soll also die in der Literatur beschriebene, hohe Falsch-Positiv-Rate der Interaktionsnetzwerke (vgl. Abschnitt Hintergrund) verringert werden.

Wie gut die einzelnen Pathways rekonstruiert werden können, soll anhand von Pathway-Datenbanken und PPI-Datenbanken evaluiert werden. Im Internet zugängliche PPI-Datenbanken repräsentieren eine Sammlung von PPIs, die aus wissenschaftlicher Literatur stammen. Die PPIs in den Datenbanken sind annotiert und die Annotationen sind von Experten manuell auf Korrektheit überprüft. Pathway-Datenbanken, ebenfalls größtenteils webseitig zugänglich sind biologische Datenbanken, die biochemische Reaktionswege (Pathways) und deren Einzelreaktionen beschreiben [Schaefer, 2004].

Vorgehen

Die Datenbasis für das Rekonstruieren von PPI-Pathways unter Verwendung eines großen Korpus' liefern biomedizinische Artikel in Textform, auf die über PubMed Zugriff besteht. Auf diesen

Texten wird dann die Satzerkennung mit JSBD[4] erfolgen. Im nächsten Schritt werden die Proteinentitäten in den Sätzen mit Hilfe des NER-Tools GNAT [Hakenberg et al., 2008] markiert.

Die Ursprungssätze inklusive der markierten Proteinkandidaten für eine potentielle PPI repräsentieren die Eingangsdaten für bewährte Methoden zur PPI-Extraktion auf Satzebene wie bspw. JSRE [Guiliano et al., 2006]. Es werden mehrere dieser Methoden auf den Eingangsdaten angewendet. Zur Ergebniserzeugung werden dabei verschiedene Parametereinstellungen (soweit von den Methoden bereitgestellt) untersucht. Es kommen letztendlich genau die Einstellungen zur Anwendung, die möglichst viele potentielle PPIs extrahieren, vorerst ohne Berücksichtigung auf Korrektheit der Extraktionsergebnisse. Ausgehend von den jeweiligen Ergebnissen der PPI-Extraktionen werden PPI-Netzwerke erstellt.

Diese Netzwerke werden vermutlich noch viele Knoten und Kanten enthalten, die keine PPI darstellen, sondern Falsch-Positiv-Ergebnisse repräsentieren. Deshalb werden die Netzwerke unter Verwendung von Filtermechanismen derart reduziert, dass die relevanten PPI-Pathways übrig bleiben. Das heißt, zur Selektion zuverlässiger PPI-Pathways aus den PPI-Netzwerken werden einige Filtermechanismen wie bspw. der Verwandtschaftsgrad (vgl. Ranking von Genen in Interaktionsnetzwerken [Özgür et al., 2008]) angewendet. Zusätzlich werden weitere Filtermechanismen aus dem Bereich der Analyse biologischer Netzwerke (siehe [Junker et al., 2008]) realisiert, die dann ebenfalls auf die Netzwerke angewendet werden.

Anschließend wird die Güte für die einzelnen Filterergebnisse (je Netzwerk für die gefilterten PPI-Pathways) ermittelt. Zu diesem Zweck werden die einzelnen Ergebnisse unter Verwendung des Wissens aus PPI-Datenbanken und Pathway-Datenbanken evaluiert. Insbesondere wird hierbei untersucht, wie sich Precision und Recall der gefilterten PPI-Pathways verhalten. Abschließend sollten Aussagen darüber getroffen werden können, unter welchen Konstellationen (PPI-Extraktionsmethode, Parametereinstellungen, Filtermechanismen) wie viele PPI-Pathways und mit welcher Güte aus den untersuchten biomedizinischen Texten gefunden werden konnten.

[4] Der JULIE Lab Sentence Boundary Detector (http://www.julielab.de/Resources/Software/NLP_Tools.html).

Literatur

[Chen et al., 2004] Hao Chen and Burt Sharp: Content-rich biological network constructed by mining PubMed abstracts, *BMC Bioinformatics Vol. 5, No. 1, 147*, 2004

[ColoNET, 2009] The ColoNET-Consortium: ColoNET - A Systems Biology Approach for Integrating Molecular Diagnostics and Targeted Therapy in Colorectal Cancer, *verfügbar: http://www.colonet.de/*, (Oktober, 2009)

[Giles et al., 2008] Cory B. Giles and Jonathan D. Wren: Large-scale directional relationship extraction and resolution, *BMC Bioinformatics Vol. 9, No. 9, S11*, 2008

[Guiliano et al., 2006] Claudio Giuliano, Alberto Lavelli and Lorenza Romano: Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature, *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006

[Hakenberg et al., 2008] Jörg Hakenberg, Conrad Plake, Robert Leaman, Michael Schroeder and Graciela Gonzalez: Inter-species normalization of gene mentions with GNAT, *Bioinformatics Vol. 24 ECCB 2008, pages i126-i132*, 2008

[Junker et al., 2008] Björn H. Junker and Falk Schreiber: Analysis of biological networks, *John Wiley & Sons, Inc., Hoboken, New Jersey*, 2008

[Özgür et al., 2008] Arzucan Özgür, Thuy Vu, Günes Erkan and Dragomir R. Radev: Identifying gene-disease associations using centrality on a literature mined gene-interaction network, *Bioinformatics Vol. 24 ISMB 2008, pages i277-i285*, 2008

[Pavlopoulos et al., 2008] Georgios A. Pavlopoulos, Anna-Lynn L. Wegener and Reinhard Schneider: A survey of visualization tools for biological network analysis, *Biodata mining, Vol. 1, No. 1, 12*, 2008

[Schaefer, 2004] Carl F. Schaefer: Pathway Databases, *Ann N Y Acad Sci, Vol. 1020, 77-91*, 2004

[Skusa et al., 2005] Andre Skusa, Alexander Rüegg and Jacob Köhler: Extraction of biological interaction networks from scientific literature, *Briefings in Bioinformatics 2005, 263-276*, 2005

[Yao et al., 2004] Daming Yao, Jingbo Wang, Yanmei Lu, Nathan Noble, Huandong Sun, Xiaoyan Zhu, Nan Lin, Donald G. Payan, Ming Li, Kunbin Qu: PathwayFinder: Paving The Way Towards Automatic Pathway Extraction, *Proc. of the Second Conference on Asia-Pacific Bioinformatics*, 2004

[Yuryev et al., 2006] Anton Yuryev, Zufar Mulyukov, Ekaterina Kotelnikova, Sergei Maslov, Sergei Egorov, Alexander Nikitin, Nikolai Daraselia and Ilya Mazo: Automatic pathway building in biological association networks, *BMC Bioinformatics, Vol. 7*, 2006