Entwicklung einer Schnittstelle zur Abfrage und algorithmischen Analyse von Protein-Protein-Interaktionsgraphen

Exposé

Benjamin Gehrels

8. März 2011

1 Motivation

Proteine und deren Interaktionen sind von großem Interesse in der biowissenschaftlichen Forschung. Es gibt daher bereits einige Datenbanken mit Informationen, welche Proteine mit welchen anderen Proteinen interagieren. (vgl. [Kim et al., 2008]) Gleichwohl liegt ein nicht unerheblicher Anteil dieser Informationen nicht in strukturierten Datenbanken, sondern in unstrukturierter Textform als Publikation in wissenschaftlichen Journalen und anderen Medien vor [Özgür et al., 2008].

Die Erschließung dieser bisher unstrukturierten Informationen für eine automatisierte Analyse und Verarbeitung und die Integration mit bestehenden, strukturierten Informationen stellt daher einen interessanten Weg bei der Erforschung biologischer Prozesse dar, beispielsweise zur Vorhersage bisher unbekannter Interaktionen [Hoffmann et al., 2005].

2 Hintergrund

Mehrere Arbeiten haben bereits versucht, aus Texten sogenannte biologische Pathways zu extrahieren. Biologische Pathways sind Graphen, die das Zusammenspiel verschiedener biologischer Entitäten (Proteine, Enzyme, etc.) beschreiben. Hierbei wurden verschiedene Methoden des Text-Minings eingesetzt, angefangen bei reinen Co-Occurrences (also dem Vorkommen von zwei Proteinnamen in einem Satz, Absatz o.ä.; beispielsweise [Stapley und Benoit, 1999]) über schlüsselwortbasierte Verfahren (beispielsweise [Blaschke et al., 1999] bis hin zu syntax- und semantikbasierten Methoden des Natural Language Processings (beispielsweise [Rindflesch et al., 1999]). Auch die Reichhaltigkeit der extrahierten Daten unterscheidet sich je nach Arbeit. So extrahieren einige nur die Tatsache, das zwei Proteine interagieren, andere hingegen extrahieren auch Beziehungen zwischen anderen biologischen Entitäten, die Art der Beziehung (beispielsweise Aktivie-

rung, Phosporylisierung) oder eine etwaige Richtung (A aktiviert B oder B aktiviert A) (vgl. [Hoffmann et al., 2005]).

Am Lehrstuhl für Wissensmanagement in der Bioinformatik der Humboldt-Universität zu Berlin wurde von Philippe Thomas durch Text-Mining in den in der PubMed Literaturdatenbank hinterlegten Abstracts biowissenschaftlicher Texte - mithin mehrerer Millionen - eine PPI[1]-Datenbank erstellt. Der Extraktionsprozess ist im Wesentlichen mit dem in [Brosy, 2010, Kap. 3] beschriebenen vergleichbar. Die so erstellte Datenbank umfasst ungefähr 170.000 Proteinnamen mit über 1,3 Millionen unterschiedlichen Interaktionen in circa 6 Millionen Sätzen aus circa 1,4 Millionen Abstracts.

3 Ziel der Arbeit

Ziel der Arbeit soll es sein, die am Lehrstuhl bereits vorliegenden PPI-Daten in ein einfach verteilbares Format zu überführen. Desweiteren soll eine API entwickelt werden, mit Hilfe derer es möglich ist, aus Software heraus PPIs zu nach Kriterien zu selektieren und Teilgraphen, die diese PPIs enthalten, abzufragen. Als Kriterien sind hier beispielsweise assoziierte Krankheiten, Wirkstoffe, Spezies oder Gene möglich. Als Nutzer dieser API soll sodann ein Plugin für die Graphenanalyse- und Darstellungssoftware Cytoscape [Shannon et al., 2003] entwickelt werden.

Hierdurch wird Nutzern eine einfache Navigation durch die enormen Mengen wissenschaftlicher Texte in PubMed ermöglicht. [Hoffmann und Valencia, 2005, 3.1] haben die These aufgestellt, das es, will man diese Texte in ein navigierbares Netzwerk konvertieren, nötig sei, diese „in Cluster gleicher Größe und gleichen Informationsgehalts zu teilen“. Hierfür haben Sie Gen- und Proteinnamen als Abgrenzungskriterium genutzt, um dann alle Sätze, die ein Gen bzw. Protein beinhalten, als einen Cluster (in Form einer Webseite) anzuzeigen und zwischen ihnen per Hyperlinks zu navigieren. Zur Darstellung in Cytoscape sollen die Textausschnitte nun auf der Ebene von Proteinpaaren (PPIs) geclustert werden. Die Textquellen sollen dann als Kantenatribute des Protein-Netzwerks dargestellt werden.

Als weiterer unmittelbarer Nutzer der Daten ist die Software „Network Curator“ [2] angedacht, welche momentan im Rahmen des ColoNet-Projektes an Institut für Biologie der HU Berlin betrieben wird. Hierbei handelt es sich um eine in Perl geschriebene Web-Anwendung, welche es erlaubt, Interaktionsgraphen zu editieren.

4 Vorgehensweise

Die Entwicklung soll mehrstufig erfolgen. Als ersten Schritt plane ich, die API auf ein einziges Selektionskriterium zu beschränken: Gegeben sei ein Protein, mit welchen anderen Proteinen interagiert dieses Protein? Hierzu werden vier Komponenten entwickelt:

[1] Protein-Protein-Interaktion

[2] `http://cheetah.biologie.hu-berlin.de/curas72/`

4.1 Datenextraktion und Schematransformation

Eine erste Komponente extrahiert die benötigten Daten aus der Quelldatenbank und transformiert diese in ein Schema, welches auf die Abfragemöglichkeiten der API hin optimiert ist. Hierbei sollen die Daten unter anderem um später nicht abfragbare Informationen bereinigt werden.

4.2 Befüllung und Paketierung der Zieldatenbank

Eine zweite Komponente erstellt dann eine Embedded-Datenbank, legt darin das zu definierende Zielschema an und befüllt dieses mit den zuvor aufbereiteten Daten aus der Quelldatenbank. Sodann wird diese Datenbank in ein einfach verteilbares Format paketiert. Dies kann beispielsweise ein JAR- oder ZIP-Archiv sein. Um die Daten auch für den „Network Curator“ nutzbar zu machen, sollen hier mehrere Möglichkeiten evaluiert werden. So wäre es beispielsweise möglich, eine Embedded-Datenbank zu nutzen, welche auch aus Perl-Programmen heraus lesbar ist. Alternativ könnte man prüfen, ob ein MySQL-Server (in einem solchen werden momentan die Daten des „Network Curator“ gehalten) auch Embedded-MySQL-Datenbanken integrieren kann. Ebenfalls wäre die Variante zu prüfen, die Daten zusätzlich als MySQL-Dump im Textformat zur Verfügung zu stellen.

4.3 Entwicklung der Java API

Eine dritte Komponente stellt die Java API dar. Diese soll nun auf die paketierte Datenbank zugreifen und die Daten als Java-Objekte zur Verfügung stellen. Die API liefert zu jeder PPI die jeweilige Quellenangabe (Corpus, Dokumenten-ID, Fundstelle im Dokument) sowie den jeweiligen Satz, aus dem die PPI extrahiert wurde. Wie oben beschrieben soll als erster Schritt eine Beschränkung auf die Abfrage eines einzelnen Proteins und dessen Interaktionen ermöglicht werden.

Die API soll zusammen mit der Embedded-Datenbank paketiert werden und somit möglichst einfach, beispielsweise als Download-Möglichkeit auf einer Website, verbreitet werden können.

4.4 Anbindung an Cytoscape per Plugin

Als vierte Komponenten soll ein Plugin für das Graphenanalyse- und Darstellungsprogramm Cytoscape entwickelt werden. Hier soll es im ersten Schritt möglich sein, durch Eingabe eines Proteinnamens die dazugehörigen PPIs über die API abzufragen und diese im aktuellen Netzwerk darzustellen.

4.5 Weitere Entwicklung

Ein erster Schritt in der weiteren Entwicklung wird sein, die bis dato existierenden Abfragemöglichkeiten so zu erweitern, dass es möglich wird, nicht nur die PPIs eines Proteins, sondern auch die PPIs der mit dem gegebenen Protein interagierenden Proteine

abzufragen. Somit wird ein Teilgraph - bis zu einer vom Benutzer wählbaren Tiefe - vom gegebenen Protein aus selektierbar.

Desweiteren denkbar wäre hier die Möglichkeit, ausgehend von einer Menge von gegebenen Proteinen (statt einem einzelnen) einen umschließenden Subgraphen zu selektieren.

Auch die Selektion anhand anderer Kriterien als dem Namen des Proteins sind denkbar. In [He et al., 2010] ist im Rahmen des „BSQA Relation Mining Subsystem“ eine solche Selektionsmöglichkeit auf Basis von Genen beschrieben. Hier werden beispielsweise Abfragen wie „Finde alle Gene, die mit dem Verhalten X in Zusammenhang stehen“ oder „Welche Gene werden von Gen X reguliert“ genannt.

Um vergleichbare Abfragen auch mit der im Rahmen dieser Arbeit zu entwickelnden API zu ermöglichen, ist die Integration von Annotationen vorgesehen. Annotationen können zum Beispiel Daten über Krankheiten und Medikamente sein (vgl 3), welche auch bereits in der Datenbasis am Institut vorliegen. Auch eine Selektion nach dem Konfidenzwert, mit dem die Interaktion aus dem Text bestimmt wurde, und Metadaten des Textes, in dem die Interaktion gefunden wurde, soll ermöglicht werden. Denkbar wären hier beispielsweise der Name des Quellmediums (Journal, o.ä.), der Name des Autors oder das Erscheinungsjahr. Beispielhaft sei hier als Anfrage gegeben: „Gegeben der Wirkstoff ‚Acetylsalicylsäure‘. Welche Protein-Protein-Interaktionen werden hierdurch beeinflusst? Suche einen Teilgraph der Tiefe 3“.

Bei Anfragen mit mehreren Selektionskriterien soll hier eine boolsche UND-Verknüpfung stattfinden. Eine weitere Verknüpfung mit ODER und NICHT-Semantik bzw. einer Kombination aus diesen soll geprüft werden.

Als zweiter Schritt ist vorgesehen, dass die API die Möglichkeit bietet, Algorithmen auf Interaktionsgraphen anzuwenden und die Graphen sodann mit den Ergebnissen zu annotieren. So könnten beispielsweise kürzeste Wege in den Interaktionsgraphen oder Zentralitätsmaße berechnet werden. Mithilfe solcher Zentralitätsmaße haben [Özgür et al., 2008] bereits Ansätze evaluiert, aus Geninteraktionsnetzwerken Assoziationen zu Krankheiten vorherzusagen (am Beispiel von Prostatakrebs). Die Algorithmen werden dabei von André Koschmieder entwickelt. Hier sollen verschiedene Möglichkeiten evaluiert werden, die Algorithmen über eine möglichst generische Schnittstelle anzubinden, so dass auch zukünftige Anforderungen und weitere Algorithmen leicht eingebunden werden können.

4.6 Dokumentation

Das Ergebnis des Entwicklungsprozesses wird im Rahmen einer Studienarbeit dokumentiert. Hierbei soll eine Einführung in das biowissenschaftliche Themengebiet den Anfang bilden. Hierauf aufbauend wird erläutert, woher die Quelldaten stammen und mit welchen Methoden sie gewonnen wurden. Bezüglich der Entwicklung werden die wesentlichen Design- und Technologieentscheidungen sowie ihre Entscheidungsgründe erläutert. Desweiteren soll auf die aus dem iterativen Vorgehensmodell gewonnenen Erfahrungen eingegangen werden. Die Arbeit soll dann mit einem Ausblick auf mögliche Erweiterungs- und Weiterentwicklungsmöglichkeiten abgeschlossen werden.

Literatur

[Altman et al., 1999] Altman, R. B., Lauderdale, K., Hunter, L., und Klein, T. E., editors (1999). *Pacific Symposium on Biocomputing 2000: Honolulu, Hawaii, USA 4-9 January 2000*. World Scientific. Online, `http://psb.stanford.edu/psb-online/proceedings/psb00/`, abgrufen am 18. Februar 2011.

[Blaschke et al., 1999] Blaschke, C., Andrade, M. A., Ouzounis, C., und Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein-interactions. In Lengauer, T., Schneider, R., Bork, P., Brutlag, D., Glasgow, J., Mewes, H.-W., und Zimmer, R., editors, *Proceedings of the seventh international conference on intelligent systems for molecular biology*, pages 60–67. International Society for Computational Biology, AAAI Press.

[Brosy, 2010] Brosy, F. (2010). Rekonstruktion biologischer Pathways aus sehr großen Korpora. Diplomarbeit an der Humboldt-Universität zu Berlin.

[He et al., 2010] He, X., Li, Y., Khetani, R., Sanders, B., Lu, Y., Ling, X., Zhai, C., und Schatz, B. (2010). BSQA: integrated text mining using entity relation semantics extracted from biological literature of insects. *Nucleic Acids Research*, 38(Supplement 2):W175–W181.

[Hoffmann et al., 2005] Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C., und Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. *Science's STKE*, 2005(283). pe21.

[Hoffmann und Valencia, 2005] Hoffmann, R. und Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21(Supplement 2):ii252–ii258. ECCB/JBI'05 proceedings. Fourth European Conference on Computational Biology/Sixth Meeting of the Spanish Bioinformatics Network (Jornadas de BioInformática), Palacio de Congresos, Madrid, Spain, September 28-October 1, 2005.

[Kim et al., 2008] Kim, S., Shin, S.-Y., Lee, I.-H., Kim, S.-J., Sriram, R., und Zhang, B.-T. (2008). PIE: an online prediction system for protein–protein interactions from text. *Nucleic Acids Research*, 36(Supplement 2):W411–W415.

[Rindflesch et al., 1999] Rindflesch, T. C., Tanabe, L., Weinstein, J. N., und Hunter, L. (1999). EDGAR: Extraction of drugs, genes and relations from the biomedical literature. In [Altman et al., 1999], pages 514–525. Online, `http://psb.stanford.edu/psb-online/proceedings/psb00/`, abgrufen am 18. Februar 2011.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., und Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504.

[Stapley und Benoit, 1999] Stapley, B. J. und Benoit, G. (1999). Biobibliometrics: Information Retrieval and Visualization from Co-Occurrences of Gene Names in Medline Abstracts. In [Altman et al., 1999], pages 526–537. Online, `http://psb.stanford.edu/psb-online/proceedings/psb00/`, abgrufen am 18. Februar 2011.

[Özgür et al., 2008] Özgür, A., Vu, T., Erkan, G., und Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285. ISMB 2008 conference proceedings 19-23 July 2008, Toronto.