Exposé zur Studienarbeit: Visualisierung im Zeitverlauf von PubMed-Veröffentlichungen zu Proteinen, Genen und Protein-Protein-Interaktionen

Marten Richert*

11.01.2012

*richert@informatik.hu-berlin.de

1 Einführung

Die biowissenschaftliche Forschung erzeugt kontinuierlich Wissen über biologische Zusammenhänge in Organismen und befasst sich dabei auf molekularer Ebene mit Genen und Proteinen. Dieses Wissen liegt zum größten Teil in Form natürlichsprachlicher Texte vor. Über das Internet oder allgemein mittels digitaler Vernetzung ist es möglich in kürzester Zeit darauf zuzugreifen. Beispielsweise sind in der öffentlich zugänglichen Datenbank PubMed des MEDLINE (Medical Literature Analysis and Retrieval System Online) über 19 Mio. Referenzen zu Veröffentlichungen mit lebenswissenschaftlichem[1] Inhalt erfasst. In 2010 kamen knapp 700000 Referenzen hinzu. Allein die Statistik dieser Sammlung [MFS11] demonstriert den großen Umfang und den rapiden Zuwachs an Information auf dem Gebiet der Biowissenschaften. Das Wachstum verläuft bisher sogar exponentiell [HC06].

Eine prominente Aufgabe der Bioinformatik besteht darin, unter Verwendung computergestützter Analyseverfahren, textübergreifende Zusammenhänge in diesen Veröffentlichungen zu erkennen. Dazu werden Dokumentensammlungen mittels *Text Mining* und Informationsextraktion (*Information Extraction*) nach enthaltenen Fakten durchsucht, die im Anschluss in strukturierte Daten überführt werden können [FZKH05]. Insbesondere ist es möglich, die Dokumente hinsichtlich in ihnen enthaltenen Entitäten, darunter Gene und Proteine, sowie Interaktionen zwischen Proteinen automatisiert zu annotieren. Mit diesen Methoden lassen sich sogar ganze Netzwerke von biochemischen Reaktionen und Reaktionssequenzen in lebenden Zellen, sogenannte metabolische Pathways, über Dokumentgrenzen hinweg erkennen (siehe Abschnitt 2.2).

Für die Untersuchung der erzeugten strukturierten Informationen spielt die Darstellung der extrahierten Entitäten und Beziehungen in grafischer Form eine wichtige Rolle. Besonders interessant ist die Betrachtung der Entwicklung des Wissens über einen bestimmten Zusammenhang im Verlauf der Zeit. Der zeitliche Bezug kann dazu primär über das Veröffentlichungsdatum einer Publikation hergestellt werden. Die grafische Darstellung von Informationen im Zeitverlauf kann helfen, komplexe Zusammenhänge und historische Entwicklungen intuitiv zu erkennen.

Die grafische Aufbereitung komplexer Beziehungsnetzwerke zwischen Proteinen dient dazu, größere Zusammenhänge als die paarweisen Beziehungen in den Netzwerken zu erfassen. Nebenbei können bei der Darstellung von Netzwerken auch ästhetische Aspekte von Interesse sein.

2 Hintergrund

Die für diese Arbeit relevanten Daten sind die aus Publikationen extrahierten Gen- bzw. Proteinnamen, genauer deren eindeutige Zuordnung zu einem Eintrag in einer Gendatenbank. Weiterhin interessieren Protein-Protein-Interaktionen (PPI), die in den Abstracts und gegebenenfalls in den Volltexten beschrieben werden.

[1] Der Begriff *Lebenswissenschaften* beinhaltet neben den zahlreichen, ursprünglich biologischen Disziplinen der *Biowissenschaften* (Biochemie, Bioinformatik, Biophysik, Botanik, Cytologie, Genetik, Histologie, Immunbiologie, Mikrobiologie, Mykologie, Neurobiologie, Ökologie, Verhaltensforschung, Zoologie, Medizin, Biomedizin, Molekularbiologie, Pharmazie oder Biodiversitätsforschung) noch weitere Wissenschaftsbereiche. Zu den *Lebenswissenschaften* können in Anlehnung an die englische Bezeichnung *Life Sciences* auch die Psychologie oder sogar Forschung über künstliche Intelligenz hinzugezählt werden. Quelle: de.wikipedia.org #Biowissenschaften und #Naturwissenschaften

2.1 Datenherkunft: Informationsextraktion

Um diese Informationen zu gewinnen, ist in beiden Fällen zunächst das Erkennen der gesuchten Entitäten in den Texten Voraussetzung. Diesen Vorgang nennt man *Named Entity Recognition* (NER). Viele NER-Tools zerlegen dazu einen natürlichsprachlichen Text in seine Token. Methoden der natürlichen Sprachverarbeitung erlauben die Analyse der grammatikalischen Struktur von Sätzen. Beispielsweise ordnen POS-Tagger jedem Token seine Wortart zu. Der Überblick über NER in [LH05] nimmt eine Unterteilung der NER-Verfahren in regelbasiert, klassifikationsbasiert, sequenzanalysebasiert und wörterbuchbasiert vor. Aufgrund ihrer guten Ergebnisse finden hauptsächlich klassifikationsbasierte Verfahren in aktuellen Tools Verwendung. Dabei spielt die Wahl des Featuresets eine wichtige Rolle, welche in [LG+08] genauer untersucht wird.

Der Einsatz von NER-Verfahren ist nötig, da die Autoren ihre Erkenntnisse in der Regel unter Verwendung von verschiedenen synonymen Namen der Entitäten dokumentieren. Es gibt zwar beispielsweise für die menschlichen Gene eine einheitliche Nomenklatur, die Human Genome Organisation (HUGO) (siehe [WLD+04]), doch Tamames und Valencia stellen in ihrer Analyse [TV06] fest, dass die Wissenschaftsgemeinde die Richtlinien von HUGO derzeit nicht ausreichend umsetzt. Und selbst wenn der HUGO-Name eines Gens im Abstract angegeben wird, so werden funktionale Beziehungen von Genen und Proteinen oftmals doch wieder unter Verwendung von Synonymen formuliert. Außerdem existiert HUGO erst seit dem Jahr 2000, so dass ältere Veröffentlichungen unmöglich auf diese Nomenklatur zurückgreifen konnten.

Es gibt auch den umgekehrten Fall, dass ein Name mehreren Genen oder Proteinen zugeordnet werden kann [FZKH05]. Auch dieser Herausforderung stellen sich NER-Tools. So ist es möglich, den Kontext, in dem ein Genname gefunden wird, mit in die Bewertung einfließen zu lassen. Dazu gehört das Erkennen der Spezies [STL10] oder die Verwendung von Hintergrundwissen z. B. in Form von Gen-Ontologie-Datenbanken [HPR+08]. Die Effektivität von NER lässt sich dadurch signifikant steigern [HPR+08].

Die Abbildung eines Gennamens auf einen eindeutigen Identifikator ist das Ergebnis der *Named Entity Normalization* (NEN). Das NEN-System GNAT [HPL+08] beispielsweise ordnet die gefundenen Gennamen den Einträgen in der Datenbank EntrezGene zu.

2.2 Verbindungen zwischen Entitäten: PPI und Pathways

Wie die Gene und Proteine an sich, können auch Beziehungen zwischen ihnen mittel Informationsextraktion aus Veröffentlichungen gewonnen werden. Der Überblick in [ZDFYC07] enthält einen Abschnitt über Methoden zur Identifizierung von Beziehungen zwischen biomedizinischen Entitäten. Beziehungen zwischen Proteinen nennt man Protein-Protein-Interaktionen (PPI). Im einfachsten Fall genügt das Erkennen von zwei Proteinen im selben Satz (*Co-Occurance*). Der Recall ist dabei hoch, die Precision hingegen sehr niedrig. Zu Verbesserung des Recalls und des F-Measures kommen NLP-Methoden (*natural language processing*) zum Taggen und Parsen zum Einsatz. Außerdem werden klassifikationsbasierte Verfahren eingesetzt. Betrachtet man mehrere Proteine, zwischen denen Interaktionen auftreten, zugleich, so lassen diese sich zu Interaktionsnetzwerken zusammenfassen. Solche Netzwerke bieten einen fundamentalen Blick auf biologische Funktionen und Prozesse [Sch04]. Die Netzwerke werden als Pathways bezeichnet.

2.3 Dimension Zeit

Für die Darstellung von zeitlicher Entwicklung in einem Diagramm ist es notwendig, dass die darzustellenden Informationen einen Zeitstempel aufweisen. Gleiches gilt für die Animation von Entwicklungen auf methabolischen Pathways. Eine Animation macht nur dann Sinn, wenn sich bestimmte Eigenschaften der dargestellten Elemente im Zeitverlauf ändern. Eine Zuordnung einer extrahierten Information zu einem Zeitpunkt ist durch das Veröffentlichkeitsdatum der Quelldokumente gegeben. Somit kann zu jedem Gen/Protein oder jeder PPI die Anzahl ihres Auftretens pro Zeitabschnitt bestimmt werden.

3 Ziel

Ziel der Arbeit ist eine Anwendung, die zwei visuelle Darstellungen des Zeitverlaufs der extrahierten Daten erzeugen kann.

Zum einen wird nach Eingabe von Genen/Proteinen ein Zeitdiagramm generiert, dass für jedes Gen/Protein die Häufigkeit des Auftretens in einem Zeitabschnitt (z. B. pro Jahr) darstellt. Auf diese Weise kann die Entwicklung der Forschung zu verschiedenen Genen/Proteinen betrachtet werden. Eine solches Diagramm kann auch für PPIs erzeugt werden. In der Literatur wird schon die Anwendung MLTrends beschrieben, mit der man nach Eingabe von beliebigen Suchphrasen solche Diagramme erzeugen kann [PAN10]. Der Unterschied zum hier verwendeten Ansatz wird im folgenden Abschnitt 4 aufgezeigt.

Die zweite vorgeschlagene Möglichkeit, einen Zeitverlauf darzustellen, ist eine Animation. Die Anwendung soll zu Pathways dynamische Grafen erzeugen, in denen die historische Entwicklung von Proteinnetzwerken verfolgt werden kann.

4 Verwandte Arbeiten

Das PatentMiner System [LAS97] ermöglicht es dem Nutzer durch Eingabe von Schlüsselwörtern oder Phrasen, Teilmengen von Dokumenten in einer Datenbank mit allen in den USA zugelassenen Patenten zu selektieren. Die Suche geschieht auf dem Titel und dem Abstract der Dokumente. Die Datenbank enthält zu jedem Patent einen Zeitstempel. Nach Einschränkung eines Zeitraumes wird dem Nutzer eine Grafik mit der Anzahl von Patenten für jedes Jahr präsentiert. Eine Phrase wird dabei mittels sequentiellem Patternmining [SA96] identifiziert. Dazu gibt der Nutzer mit einer definierten Syntax eine Anfrage ein, die mehrere Schlüsselwörter enthalten kann. Das erzeugte Diagram stellt Trends vergleichbar mit dem ersten Ziel dieser Arbeit dar. Grundlage bilden jedoch die eingegebenen Anfragesequenzen im Gegensatz zu der in dieser Arbeit vorgesehenen Möglichkeit, erkannte Entitäten darzustellen. Weiterhin ist es mit PatentMiner möglich, nach Trendverläufen wie Ab-/Aufwärtstrend, Spitzen oder Wiederanstieg für die Anfragen zu suchen. Dazu werden *Shape queries* [APWZ95] verwendet. Eine derartige Trendsuchmöglichkeit ist in dieser Arbeit nicht vorgesehen.

Mit der Webapplikation MLTrends lassen sich Häufigkeiten von Termen oder Phrasen in biowissenschaftlich Veröffentlichungen über den Verlauf der Zeit grafisch veranschaulichen. Der Nutzer gibt dabei feste Suchstrings ggf. logisch verknüpft in eine Suchmaske ein. MLTrends gibt nun für

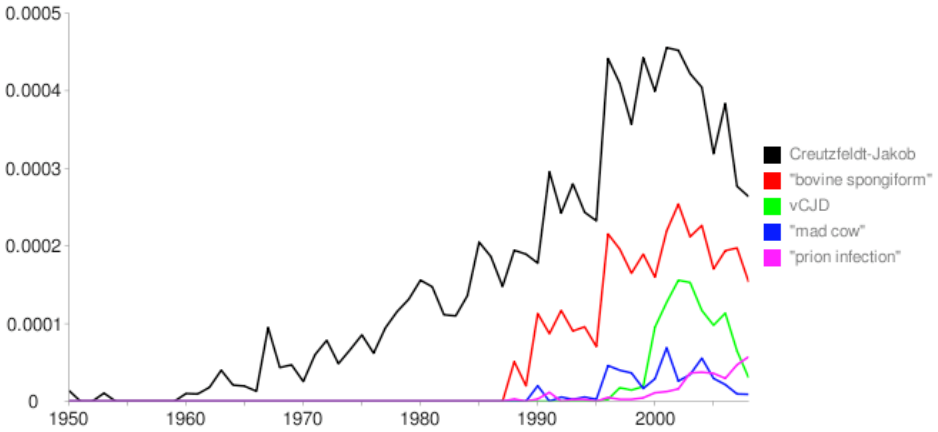

Abbildung 1: Die Ausgabe von MLTrends zeigt die Verteilung der MEDLINE-Artikel, die zu gegeben Suchtermen gefunden wurden, nach Publikationsjahr. Die Dokumentenzahl in jedem Datenpunkt ist hier im Verhältnis zur Gesamtanzahl aller Dokumente eines jeden Jahres angeben. (Quelle der Abbildung: [PAN10] Seite 2)

jedes Jahr die Anzahlen der auf die Suchanfragen zutreffenden Dokumente in Form eines Diagramms aus. Dazu indiziert MLTrends alle Abstracts und Titel aus MEDLINE für eine schnelle Volltextsuche auf diesen Daten. Das Ergebnis der Volltextsuche bestimmt für jedes Jahr die Anzahlen der Dokumente, auf die die Suchstrings zutreffen, [PAN10]. Somit ist es beispielsweise möglich, die Verwendung von Gen- oder Proteinnamen über den Zeitraum der dokumentierten biowissenschaftlichen Forschung visuell zu veranschaulichen. Allerdings können dabei nur exakte Treffer zu den Suchanfragen gezählt werden. Abbildung 1 zeigt ein mit MLTrends erzeugtes Diagram als Resultat einer Suche nach mehreren verschiedenen Termen. Eine ähnliche Verlaufsdarstellung ist auch Ziel dieser Arbeit. Jedoch im Unterschied zu MLTrends ist das Kriterium, welches entscheidet, ob ein Dokument gezählt wird, nicht das exakte Finden einer Zeichenfolge, sondern die durch Text Mining und Informationsextraktion erzeugte Verknüpfung zwischen Dokument und Gen. Beziehungen zwischen Entitäten oder Netzwerke in ihrer Entwicklung lassen sich nicht darstellen. Das ist jedoch ein Ziel dieser Arbeit.

In [LS08] werden interdisziplinäre Entwicklungen in der Wissenschaft mit Hilfe von *Journal Citation Reports* (JCR) untersucht. Dabei finden auch visuelle Methoden Verwendung. Es werden dynamische Grafen erzeugt, an Hand derer man den Zitations-Einfluss (citation-impact) auf die Journale *Cognitive Science*, *Social Networks* und *Nanotechnology* im Verlauf von Jahren verfolgen kann. Interessant ist hier die Darstellung der Knoteneigenschaft *Betweenness centrality* mittels der Größe der Knotenpunkte. Diese Größe kann sich während des Zeitablaufs verändern, was in der erzeugten Grafenanimation gut erkennbar ist.

5 Inhalt der Studienarbeit

5.1 Umfang

Geplant ist die Erstellung einer Anwendung zur visuellen Analyse der Häufigkeiten von in Dokumenten gefundenen Genen und Proteinen. Dazu wird dem Nutzer die Auswahl eines Gens oder

einer Menge von Genen aus EntrezGene oder UniProt ermöglicht. Unter Einbindung der am Lehrstuhl vorliegenden, aus MEDLINE-Abstracts extrahierten Daten über Entitäten, darunter Gene und Proteine, wird die Darstellung eines Zeitdiagramms analog zu MLTrends implementiert. Die Darstellung soll auch für Protein-Protein-Interaktionen anwendbar sein.

Eine weitere Möglichkeit der Visualisierung von Proteininteraktionen ist die Ausgabe einer Animation, die eine gegebene Menge von Proteinen auf einer Fläche anordnet. Die historisch dokumentierten Interaktionen werden nun durch im Zeitverlauf stärker werdende Verbindungslinien dargestellt. Möglicherweise lässt sich auch eine sinnvolle Auswertung der Interaktionshäufigkeiten für die Positionierung oder Formgebung der Knotenpunkte finden.

Es sollen dem Anwender Möglichkeiten geschaffen werden, unter den Millionen von Entitäten solche zu wählen, die in irgendeiner Form von Interesse sein könnten. Folgende Möglichkeiten seien dazu vorgeschlagen:

- Der Nutzer kann eine Liste von Gennamen eingeben.
- Der Nutzer kann eine Liste von UniProt-Ids eingeben.
- Für die Animation eines Netzwerks, kann der Nutzer einen Pathway aus der Pathway-Datenbank des Lehrstuhls wählen.
- Bei der NER wird auch die behandelte Spezies annotiert. Es können also Filter und Selektionsmöglichkeiten für den Parameter Spezies entwickelt werden.

Weiterhin wird ein spezielles Zeitdiagramm generiert, welches die Gesamtanzahl erstmalig erwähnter Proteininteraktionen pro Periode darstellt.

Kann der Zeitpunkt der Untersuchung einer PPI als Information für das Grafenlayout herangezogen werden? Denkbar ist die Positionierung zweier Proteine weiter voneinander entfernt, je früher ihre Interaktion beschrieben wurde. Oder umgekehrt.

5.2 Ausgangssituation

Der Lehrstuhl verfügt über auf mehrere Datenbanken verteilte Tabellen, die die extrahierten Entitäten wie Gene und Proteine enthalten sowie deren Zuordnung zu ihrem Ursprungsdokument. Weiterhin liegt eine Tabelle mit extrahierten PPIs und eine Tabelle mit metabolischen Pathways vor. Die Daten werden u. a. für die Applikation GeneView [TSJ+10] aktuell gehalten. Das Veröffentlichungsdatum und der Titel eines jeden Dokuments kann einem Lucene-Index-File entnommen werden. Weitere im Internet zugängliche Datensammlungen erlauben ein Mapping der verwendeten Gen-Ids auf den Gennamen.

5.3 Vorgehensweise

Zunächst werden alle vorhanden Daten verknüpft, so dass eine programmatisch gut handhabbare Abfrage aller benötigten Attribute möglich ist. Dazu werden Datenstrukturen entworfen, die mit entsprechenden Lookup-Methoden zu versehen sind. Wenn nötig wird eine einfache lokale Datenbank erstellt, die nur die benötigten Attribute enthält. Dies sollte Geschwindigkeitsvorteile und eine Vereinfachung der Handhabung ermöglichen.

Es wird eine geeignete Nutzerschnittstelle entwickelt, die eine sinnvolle Parametrisierbarkeit für den Nutzer ermöglicht. Die GUI soll dem Nutzer die im Abschnitt 5.1 beschriebenen Anfragemöglichkeiten bieten.

Für die Erzeugung eines Zeitdiagramms schlage ich die Google-Chart-Api vor.

Zur Erzeugung von Netzwerkgrafen mit force directed Layout kommt die Javabibliothek Prefuse in die engere Wahl. Ein erster Blick in Codebeispiele lässt den Schluss zu, dass auch das Erzeugen von Zwischengrafen für ein animiertes Netzwerk damit möglich ist. Die Grafen lassen sich in Java leicht als Einzelbilder exportieren. Aus den Einzelbildern kann mit einer freien Video-Bibliothek oder einem Tool wie MEncoder ein Video generiert werden. Andere Animationsmöglichkeiten mit aktuellen Webtechniken wie Javascript, Html5 oder SVG sind ebenfalls denkbar. Im Zuge dieser Arbeit wird nur eine Variante umgesetzt.

Die einzelnen Komponenten werden schließlich in einer Webapplikation zusammengefasst. Der Nutzer kann dann in einem Browser die Parameter seiner Anfrage eingeben und er bekommt je nach Usecase ein Zeitdiagramm oder eine Animation dargestellt. Die Animation kann pausiert oder schrittweise vorwärts und rückwärts abgespielt werden. Im Falle von Video ist abhängig vom Player u. U. nur ein Springen zu diskreten Zeitpunkten möglich.

A Links

EntrezGene	http://www.ncbi.nlm.nih.gov/sites/entrez/?db=gene
Google-Chart-Api	http://code.google.com/apis/chart/
Grafenanimation	http://www.leydesdorff.net/journals/cognsci/index.htm
MEncoder	http://www.mplayerhq.hu/DOCS/HTML/de/mencoder.html
MLTrends	http://www.ogic.ca/mltrends
Prefuse	http://prefuse.org/gallery/graphview/
PubMed	http://www.ncbi.nlm.nih.gov/pubmed
UniProt	http://www.uniprot.org
#Biowissenschaften	http://de.wikipedia.org/wiki/Biowissenschaften#Biowissenschaften.2C_Life_Sciences_und_Lebenswissenschaften
#Naturwissenschaften	http://de.wikipedia.org/wiki/Einzelwissenschaft#Naturwissenschaften

Literatur

[APWZ95] Rakesh Agrawal, Giuseppe Psaila, Edward L. Wimmers, and Mohamed Zaït. Querying shapes of histories. In *Proceedings of the 21th International Conference on Very Large Data Bases*, VLDB '95, pages 502–514, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[FZKH05] J. Fluck, M. Zimmermann, G. Kurapkat, and M. Hofmann. Information extraction technologies for the life science industry. *Drug Discovery Today: Technologies*, 2(3):217–224, 2005.

[HC06] L. Hunter and K.B. Cohen. Biomedical language processing: Perspective what's beyond pubmed? *Molecular cell*, 21(5):589, 2006.

[HPL+08] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez. Inter-species normalization of gene mentions with gnat. *Bioinformatics*, 24(16):i126, 2008.

[HPR+08] J. Hakenberg, C. Plake, L. Royer, H. Strobelt, U. Leser, and M. Schroeder. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome biology*, 9(Suppl 2):S14, 2008.

[LAS97] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. *Proceedings of the Third International Conference On Knowledge Discovery and Data Mining, pages 227-230*, 1997.

[LG+08] R. Leaman, G. Gonzalez, et al. Banner: An executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, 13:652–663, 2008.

[LH05] U. Leser and J. Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357, 2005.

[LS08] L. Leydesdorff and T. Schank. Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. *Journal of the American Society for Information Science and Technology*, 59(11):1810–1818, 2008.

[MFS11] *MEDLINE Facts Sheet*. U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894 USA. National Institutes of Health, 2004-2011.
http://www.nlm.nih.gov/pubs/factsheets/medline.html
abgerufen am 23.12.2011.

[PAN10] G.A. Palidwor and M.A. Andrade-Navarro. Mltrends: Graphing medline term usage over time. *Journal of biomedical discovery and collaboration*, 5:1, 2010.

[SA96] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology—EDBT'96*, pages 1–17, 1996.

[Sch04] C.F. Schaefer. Pathway databases. *Annals of the New York Academy of Sciences*, 1020(1):77–91, 2004.

[STL10] I. Solt, D. Tikk, and U. Leser. Species identification for gene name normalization. *BMC Bioinformatics*, 11(Suppl 5):P5, 2010.

[TSJ+10] P. Thomas, J. Starlinger, C. Jacob, I. Solt, J. Hakenberg, and U. Leser. Geneview–gene-centric ranking of biomedical text. *Gene*, 2010.

[TV06] J. Tamames and A. Valencia. The success (or not) of hugo nomenclature. *Genome Biology*, 7(5):402, 2006.

[WLD+04] H.M. Wain, M.J. Lush, F. Ducluzeau, V.K. Khodiyar, and S. Povey. Genew: the human gene nomenclature database, 2004 updates. *Nucleic acids research*, 32(suppl 1):D255, 2004.

[ZDFYC07] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K.B. Cohen. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358, 2007.