Ermittlung von Zusammenhängen zwischen enzymatischer Aktivität und Krankheiten durch die automatische Analyse wissenschaftlicher Publikationen

Inaugural - Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Oliver Hofmann

aus Köln

Köln 2004

meinen Eltern

I'm personally convinced that the whole goal of artificial programming is so that one day computers may know the frustration of trying to program themselves.

Comment on Deft Code:
Every Language War Ever

Danksagung

Bedanken möchte ich mich bei Herrn Prof. Dr. Schomburg für die Bereitstellung des Themas, seine Unterstützung in der Arbeitsgruppe und die Diskussionsbereitschaft. Ebenso möchte ich den Mitarbeitern der Arbeitsgruppe danken, allen voran Christian aus dem Spring für seine Hilfsbereitschaft, einen unvergleichlichen Humor und die Art, Dinge auf den Punkt zu bringen; Ida Schomburg, Antje Chang und Christian Hoppe für ihre freundliche Hilfe zu jeder möglichen und unmöglichen Zeit.

Ein großer Dank an Olav Zimmermann für inspirierende Diskussionen und die Bereitschaft, seine genialen Einfälle zu teilen. Ein ebenso großer Dank an Vera Grimm für Ihre nicht minder verrückten Ideen und Ihre Freundschaft.

Insbesondere möchte ich meinen Eltern für Ihre Unterstützung und Geduld danken, womit sie mir dieses Studium erst ermöglichten.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Diagrammen und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat, dass sie – abgesehen von der in Anhang A.9 angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Dietmar Schomburg betreut worden.

Oliver Hofmann

1. Referent: Prof. Dr. D. Schomburg
2. Referent: Prof. Dr. S. Waffenschmidt
Eingereicht: 8. April 2004
Disputation: 9. Juli 2004

Inhaltsverzeichnis

Abkürzungsverzeichnis

BRENDA	Braunschweiger Enzymdatenbank
EC	Enzyme Commission
GO	Gene Ontology
IUBMB	International Union of Biochemistry and Molecular Biology
KEGG	Kyoto Encyclopedia of Genes and Genomes
MESH	Medical Subject Headings
NCBI	National Center for Biotechnology Information
OMIM	Online Mendelian Inheritance in Man
PMID	PubMed Identifier
SVM	Support Vector Machine
UMLS	Unified Medical Language System

Abstract

Given the explosive growth of biomedical data as well as the literature describing results and findings, it is getting increasingly difficult to keep up to date with new information. Keeping databases synchronized with current knowledge is a time-consuming and expensive task, one which can be alleviated by automatically gathering findings from the literature using linguistic approaches. This dissertation describes a method to automatically annotate enzyme classes with disease-related information extracted from the biomedical literature. Enzyme names for the 3 901 enzyme classes in the *Braunschweiger Enzymdatenbank* (BRENDA) database, a repository for quantitative and qualitative enzyme information, were identified in more than 100 000 abstracts retrieved from the *PubMed* literature database. Phrases in the abstracts were assigned to concepts from the *Unified Medical Language System* (UMLS) utilizing the *MetaMap* program, allowing for the identification of disease related concepts by their semantic fields in the UMLS ontology. Assignments between enzyme classes and and diseases were created based on their co-occurrence within a single sentence. False positives could be removed by a variety of filters including minimum number of co-occurrences, removal of sentences containing a negation and the classification of sentences based on their semantic fields by a Support Vector Machine. Verification of the assignments with a manually annotated set of 1 500 sentences yielded favorable results of 95 % precision, sufficient for inclusion in a high-quality database.

Kurzzusammenfassung

Aufgrund des schnellen Wachstums biomedizinischer Daten sowie der assoziierten Literatur wird es auch für Experten zunehmend schwierig, den Überblick über den aktuellen Wissensstand zu behalten. Der Aufbau und die manuelle Erweiterung von Datenbanken ist teuer und zeitaufwändig, kann jedoch durch linguistische Methoden unterstützt werden, welche Erkenntnisse automatisch aus der wissenschaftlichen Literatur extrahieren. Die vorliegende Dissertation stellt eine solche Methode zur Annotation von Enzymklassen mit krankheitsrelevanten Informationen vor. Die Enzymnamen von 3 901 Enzymklassen der *Braunschweiger Enzymdatenbank* (BRENDA), einer Sammlung von qualitativen und quantitativen Enzymdaten, wurden in einem Textkorpus aus über 100 000 Kurzzusammenfassungen der *PubMed*-Datenbank identifiziert. Phrasen der Kurzzusammenfassungen konnten durch das *MetaMap*-Programm den Konzepten des *Unified Medical Language System* (UMLS) zugewiesen werden, was eine Identifikation der krankheitsrelevanten Begriffe mittels ihrer semantischen Felder in der UMLS-Ontologie erlaubte. Eine Zuordnung von Enzymklassen zu Krankheitskonzepten erfolgte aufgrund der gemeinsamen Nennung innerhalb eines Satzes. Die Zahl falscher Zuordnung konnte durch den Einsatz verschiedener Filter verringert werden. Verwendet wurden unter anderem die Mindestzahl gemeinsamer Nennungen, die Entfernung von Sätzen mit einer Negation sowie die Klassifikation unbekannter Sätze durch eine *Support Vector Machine*. Eine Überprüfung der Zuordnungen anhand 1 500 manuell annotierter Sätze ergab eine Präzision von 95 %, was eine direkte Erweiterung der BRENDA-Datenbank mit den gefundenen Zuordnungen erlaubte.

1 EINLEITUNG

> In theory, there is no difference between theory and practice. But, in practice, there is.
>
> *Jan L.A. van de Snepscheut*

Mit dem Beginn der Sequenzierung des humanen Genoms hat die Automatisierung biologischer Experimente rapide zugenommen. Mehr und mehr Daten werden analysiert und interpretiert. Verdeutlicht wird dies durch das Wachstum der *PubMed*-Literaturdatenbank[1], eine der wichtigsten Quellen für biomedizinische Publikationen. Mit über 12 Millionen Einträgen und einer steigenden Wachstumsrate von zur Zeit etwa 40 000 Publikationen pro Monat spiegelt sie die Informationsflut wieder, die von Wissenschaftlern bewältigt werden muss [26, 166]. Der größte Teil der bisher erlangten biomedizinischen Erkenntnisse liegt in Form solcher wissenschaftlichen Publikationen vor. Die in den Veröffentlichungen verwendete natürliche Sprache besitzt jedoch eine zu komplexe Struktur, um das enthaltene Wissen direkt computergestützt auswerten zu können. Hierfür sind strukturierte Daten notwendig, die eine Verknüpfung untereinander sowie eine Aufbereitung je nach Wünschen des Benutzers erlauben. In den letzten Jahren haben daher maschinenlesbare Datenbanken, insbesondere

[1] Erreichbar über `http://www.pubmed.gov/`

im Hinblick auf die Bioinformatik, zunehmend an Bedeutung gewonnen[2]. Vorhandenen Datenbanken fehlt allerdings oftmals die für eine automatische Auswertung notwendige Struktur. Ein Beispiel hierfür ist die *Online Mendelian Inheritance in Man* (OMIM)-Datenbank, die qualitativ hochwertige Informationen zu vererbbaren menschlichen Krankheiten enthält [101]. Die Einträge liegen ohne ein klar definiertes Vokabular in Textform vor, wobei historischer Abriss, Krankheitsbeschreibung und Forschungsansätze ineinander übergehen. Dies mag für einen Experten bei einer interaktiven Nutzung hinreichend sein, erschwert aber die computergestützte Bearbeitung.

Zudem ist der Aufbau von Datenbanken Aufbau und die Pflege durch die manuelle Annotation von Publikationen mit hohen Kosten verbunden. Ein Beispiel für den notwendigen Aufwand bietet die *Braunschweiger Enzymdatenbank* (BRENDA), welche quantitative und qualitative Informationen zu 3 901 Enzymklassen enthält. Daten zu mehr als 80 000 Enzymen wurden manuell aus 50 000 Publikationen gesammelt und in 40 Tabellen einer relationalen Datenbank organisiert, die eine maschinengestützte Auswertung ermöglicht [135]. Das erwähnte schnelle Wachstum der Literatur führt allerdings zu Lücken in den annotierten Daten, so dass selbst eine laufend aktualisierte Datenbank wie BRENDA in Teilbereichen hinter dem bereits publizierten Wissen zurückbleibt.

Eine Möglichkeit, solche Lücken zu füllen, ist die Erweiterung der manuell annotierten Daten mittels eines automatisch generierten Anteils. So enthält *Swiss-Prot*, eine Proteinsequenzdatenbank, manuell verifizierte Informationen zu Proteinsequenzen, während *TrEMBL* (*translated EMBL*) diese durch automatisch übersetzte Nukleinsäuresequenzen ergänzt [84].

Die für dieses Verfahren notwendigen zusätzlichen Informationen – in diesem Fall die Nukleinsäuresequenzen der EMBL-Datenbank – stehen nur für wenige andere Wissensgebiete zur Verfügung [21]. Daher wird zunehmend versucht, die in wissenschaftlichen Publikationen enthaltenen Daten direkt auswerten. Hierzu werden Systeme benötigt, welche natürliche Sprache in die benötigte Struktur umwandeln und dem Nutzer in leicht zugänglicher Form präsentieren. Die Entwicklung eines solchen Systems

[2] Eine Übersicht über mehr als 500 im *World Wide Web* erreichbare Datenbanken findet sich in [64].

zur automatischen Erweiterung der BRENDA-Enzymdatenbank war das Ziel der vorliegenden Arbeit.

Linguistik und elektronische Datenverarbeitung

Mit der automatischen Verarbeitung und dem Verständnis von Texten beschäftigt sich das Feld der Computerlinguistik seit den sechziger Jahren des letzten Jahrhunderts[3]. Sie setzt sich mit der formellen Analyse von natürlicher Sprache und der computergeeigneten Repräsentation auseinander. Bereits 1945 gab es die ersten Visionen, wie Technologie bei der Manipulation von Informationen dienlich sein könnte [28]. Die 1956 und 1957 publizierten Arbeiten von Noam Chomsky zur Entwicklung einer ‚universellen Grammatik' und seine Hierarchie der Sprachen erlaubten eine mathematische, formelle Betrachtung der Sprachen, die sich auf andere Wissensgebiete auswirkte[4] [38]. Studien von 1961 beschreiben die Notwendigkeit, Texte nach semantischen Kriterien abfragen zu können und Wissen aus unterschiedlichen Vokabularen aufeinander abzubilden – Techniken, die erst jetzt, 40 Jahre später, Einzug in Computerprogramme finden. Zur gleichen Zeit entstanden die ersten Arbeiten zum Vergleich von Dokumenten mit Hilfe einer Vektorrepräsentation, die zur Suche in großen Datenbanken wie *Medline* entwickelt wurden [127]. Gegen Ende der sechziger Jahre fanden die ersten Systeme zur Übersetzung von Texten der russischen in die englische Sprache Verwendung. Dabei wurde jedoch deutlich, dass die Schwierigkeiten der gesetzten Ziele drastisch unterschätzt worden waren [133].

Schwierigkeiten bei der automatischen Auswertung der natürlichen Sprache

Die Schwierigkeiten bei der Definition und dem Verständnis von natürlicher Sprache erstrecken sich über alle strukturellen Ebenen und fangen schon bei der anscheinend einfachen Abgrenzung an, was denn eigentlich ein Wort ist. Worte werden aus Morphemen gebildet, dem kleinsten

[3] Eine gute Übersicht zu den Anfängen der Computerlinguistik und der Motivationen findet sich in Schatz [133].

[4] Der besondere Einfluss der Linguistik auf Methoden der Biologie und Bioinformatik wird in Searls diskutiert [137].

bedeutungstragenden Element einer Sprache. Neben grammatikalischen Morphemen wie ‚aber' und ‚in' stellen lexikalische Morpheme wie ‚Protein' oder ‚früh' als selbstständige Wörter die wichtigste Klasse[5]. Aus der Bedeutung der Morpheme allein lässt sich aber nicht automatisch die Bedeutung eines Wortes ableiten. Ein übliches Beispiel hierfür ist die Definition des Schneemanns, einer aus Schnee bestehenden Figur, und der Erweiterung dieser Definition auf den Milchmann. Dies lässt sich problemlos auf einen Begriff wie ‚Aktionspotential' übertragen, welcher für den Biologen eine über seine Bestandteile hinausgehende Bedeutung hat. Die allgemeine Definition der von Leerzeichen begrenzten Zeichenfolge macht bei zusammengesetzten Wörtern wie ‚Diabetes Mellitus' wenig Sinn. Die Entfernung von Bindestrichen bereitet dagegen bei Begriffen wie ‚Kapital-Lebensversicherung' Schwierigkeiten und ist erst recht nicht sinnvoll bei einer chemischen Verbindung wie 1,4-Cyclohexadien.

Den lexikalischen Schwierigkeiten folgt auf der nächsten Sprachebene die Syntax eines Textes, also die Regeln, nach denen Wörter zu Teilsätzen – sogenannten Phrasen – oder vollständigen Sätzen zusammengefügt werden. In den von Sonderfällen durchsetzten natürlichen Sprachen hat die Syntax in der Linguistik eine beschreibende Funktion[6]. Bereits kleine Änderungen in der Beschreibung der Abhängigkeit von Worten zueinander können die Bedeutung eines Satzes völlig ändern, insbesondere in einer flexionsarmen[7] Sprache wie der Englischen:

— *Driving cars is dangerous. Driving cars* are *dangerous.*

Sätze mit mehr als einer Bedeutung können zumeist nur über ihren semantischen Kontext interpretiert und verstanden werden. Noam Choms-

[5] Neben diesen beiden Morphemklassen gibt es noch die Einteilung in freie und gebundene Morpheme, wobei letztere nur in Kombination mit anderen Morphemen vorkommen können: So haben ‚Him(beere)' und ‚Heidel(beere)' nur in Verbindung mit dem freien Morphem ‚Beere' einen Sinn. Weitere Klassen siehe http://de.wikipedia.org/wiki/Morphem.

[6] Oder anders ausgedrückt: „*Unfortunately, or luckily, no language is tyrannically consistent. All grammars leak.*" (Edward Sapir, 1921)

[7] Die Änderung oder Beugung eines Wortes mit seiner grammatikalischen Funktion wird als Flexion bezeichnet. Flexionen eines Wortes werden verwendet, um unter anderem Tempus, Kasus oder Singular und Plural anzuzeigen: *inhibit* → *inhibits* → *inhibited*...

kys berühmtes Beispiel für einen grammatikalisch völlig richtigen, aber inhaltlich sinnlosen Satz macht dies besonders deutlich [38]:

— *Colorless green ideas sleep furiously.*

Erst die semantischen Eigenschaften, also die Bedeutung von Worten und Konzepten der Sprache erlauben eine Entscheidung darüber, ob eine Idee an sich die Eigenschaft einer Farbe haben kann. Ohne ein aufwändiges semantisches Netzwerk, welches die Konzepte einem semantischen Bereich oder Feld zuordnet, und diese Felder dann zueinander in Beziehung setzt, ist eine computergestützte Auswertung kaum möglich. Aber selbst mit einem solchen Regelwerk bleiben Ambiguitäten:

— *Er sieht den Mann mit dem Fernglas.*

Hierbei ist unklar, wer von beiden Männern das Fernglas besitzt. Semantische Doppeldeutigkeiten lassen sich meist nur aufgrund des Kontexts auflösen, das heißt unter Einbeziehung der zuvor beschriebenen Ereignisse. Während einem Leser klar sein mag, dass es sich um ein am Anfang des Kapitels beschriebene Person handeln muss, ist dies für einen Computer nur über eine komplizierte Diskursanalyse nachzuvollziehen. Das gilt erst recht für das sogenannte Welt- oder Kulturwissen. Das ‚auf Holz klopfen' noch eine zusätzliche Bedeutung hat oder ‚in die Luft gehen' nicht unbedingt wörtlich gemeint ist, kann nur über die Erstellung einzelner Regeln implementiert werden. Jede Doppeldeutigkeit auf der lexikalischen, syntaktischen oder semantischen Ebene verdoppelt die Anzahl der möglichen Lesarten; bei komplexen Sätzen sind hunderte verschiedener Lesarten nicht unüblich. Dies erklärt, warum es nach wie vor kein System gibt, welches Texte von einer in die andere Sprache automatisch und fehlerfrei übersetzen kann – trotz gegenteiliger Versprechungen von Firmen, die diese Fähigkeit seit über 20 Jahren ankündigen [139].

Ansätze der Computerlinguistik

Um natürliche Sprache trotz der beschriebenen Probleme auswerten zu können, verfolgt die Computerlinguistik zwei verschiedene Ansätze: Die des regelbasierten *Natural Language Understanding* sowie die statistische

Linguistik[8]. Regelbasierte Verfahren konzentrieren sich dabei auf die formalen Definitionen der Syntax und Semantik. Diese sind dabei transparenter und für den Benutzer leichter zu verstehen als die schwerer zu interpretierenden Klassifikationen durch statistische Verfahren. Regelbasierte Ansätze bringen oft eine höhere Genauigkeit mit sich, da sie von Experten erstellt werden, die das jeweilige Feld, das verwendete Vokabular und die gängigen Ausdrucksweisen genau kennen. Der größte Nachteil ist der manuelle Aufwand, der zu ihrer Erstellung und Anpassung an neue Themengebiete nötig ist. Statistische Verfahren dagegen arbeiten mit Wahrscheinlichkeiten, die sie durch die Analyse existierender, annotierter Textsammlungen – sogenannter Textkorpora – lernen. Dies reicht von der einfachen Analyse von Wortfrequenzen über das Lernen von Kollokationen (das gemeinsame Auftreten zweier Wörter) zu intelligenteren Systemen, die mittels Regeln die zu untersuchenden Texte präprozessieren. Die beiden wichtigsten Hilfsmittel hierzu sind die Analyse der Syntax durch entsprechende Parser[9] sowie die Verwendung von Lexika und Ontologien[10] zum Verständnis des verwendeten Vokabulars. Die Syntax alleine reicht nicht, um die im folgenden Satz enthaltene Information auszuwerten[11]:

> — *Voltage-gated sodium and potassium channels are involved in...*

Ein Parser könnte hierbei die Komponenten *voltage-gated sodium* zusammenfassen, ohne zu erkennen, dass nicht das Natrium, sondern die Kanäle durch die Spannung verändert werden. Ebenso ist unklar, ob es sich

[8] Die Lehrbücher *Foundations of Statistical Natural Language Processing* und *Natural Language Understanding* geben eine gute Übersicht über die Methoden und Möglichkeiten der Computerlinguistik [1, 98]. Für ein deutlich kürzere Einführung siehe [123].

[9] Syntaxparser analysieren den Aufbau von Sätzen aus Substantiv- und Verbphrasen sowie einer Vielzahl an Modifikatoren wie Adjektiven oder Adverbien [134]. Die dabei verwendeten Lexika zur Identifikation von Worten enthalten zusätzliche Informationen über deren Eigenschaften und ihre möglichen Beziehungen untereinander, wodurch eine recht zuverlässige Syntaxanalyse einfacher Sätze möglich wird.

[10] Bei einer Ontologie handelt es sich um ein formales System von einheitlich definierten Konzepten, deren Bezeichnungen sowie semantischen Relationen zwischen den Konzepten.

[11] Das Beispiel wurde der *BioNLP*-Webseite entnommen: `http://www.bionlp.org/`

hier um zwei getrennte Kanäle für Natrium und Kalium handelt, oder ob es nur einen gemeinsamen Kanal gibt. Erst die in einer Ontologie enthaltenen semantischen Beziehungen weisen darauf hin, dass chemische Elemente nicht durch elektrische Spannung geöffnet werden.

Ontologien in der Biomedizin

Neben den semantischen Informationen ist eine Grundidee biologischer Ontologien, die teilweise sehr unterschiedlichen Definitionen selbst grundlegender Konzepte in verschiedenen Datenbanken zu vereinheitlichen und so einen Datenaustausch zu ermöglichen. Beispielsweise definiert *The Genome Database* ein Gen als „ein DNA-Fragment, das transkribiert und in ein Protein translatiert werden kann“, während *Genbank* den gleichen Begriff mit „einer benannten DNA-Region von biologischem Interesse und verantwortlich für ein genetisches Merkmal oder einen Phänotyp“ beschreibt [136]. Die letzte Definition beinhaltet im Gegensatz zur ersten genetische Strukturmerkmale wie Promotoren, Enhancer und Introns.

Neben allgemeinen Ontologien wie *Wordnet* und spezialisierten Lexika wie ECOCYC [85] werden in der Praxis hauptsächlich zwei große Ontologien verwendet: die *Gene Ontology* (GO) und das *Unified Medical Language System* (UMLS). Die GO-Ontologie legt ihren Schwerpunkt auf die Entwicklung eines Vokabulars zur Beschreibung der Rolle von Genen, ihrer Funktionsweise und den zellulären Komponenten von Organismen, um diese Konzepte datenbankübergreifend miteinander verbinden zu können [68]. Informationen zu Krankheiten und Syndromen sind in der GO-Ontologie so gut wie nicht enthalten.

Die seit 1986 existierende UMLS-Ontologie hat dagegen einen Schwerpunkt im medizinischen Bereich und beinhaltet drei Komponenten [106]: die wichtigste ist der sogenannte Metathesaurus, in dem Bezeichnungen zu Konzepten gruppiert vorliegen. Konzepte des Metathesaurus sind miteinander durch neun verschiedene Beziehungstypen verknüpft, welche den ursprünglichen Vokabularen entnommen sind und in den meisten Fällen eine Hierarchie repräsentieren. Als zweite Komponente beinhaltet die Ontologie ein semantisches Netzwerk allgemeinen Wissens über die Konzepte sowie ihre Beziehungen untereinander. Einzelne Konzepte sind einem oder mehreren von über 140 semantischen Feldern wie ‚Labormethode‘ oder ‚qualitatives Konzept‘ zugewiesen. Als dritte Komponente lie-

fert das SPECIALIST-Lexikon lexikalische Informationen zu Einträgen des Metathesaurus. Unter anderem enthält es die unterschiedlichen Flexionen der den Konzepten entsprechenden Bezeichnungen [17].

Anwendungen der Computerlinguistik

Die Computerlinguistik findet hauptsächlich in zwei Bereichen ihre Anwendung: zum einen dienen die Methoden als Schnittstelle zu Programmen, zum Beispiel in der Spracherkennung, der Umwandlung von Anfragen in eine formale Datenbankabfrage oder der automatischen Rechtschreibekorrektur einer Textverarbeitung. Zum anderen sind die linguistischen Methoden die eigentliche Hauptanwendung, wie bei der Informationsextraktion oder den untereinander verwandten Methoden des Textvergleichs, der Informationsfilterung und der Klassifizierung von Dokumenten. Im Idealfall wird die Nutzung linguistischer Verfahren durch den Benutzer nicht wahrgenommen – nur wenige Menschen machen sich bei der Benutzung von Suchmaschinen für das *World Wide Web* oder der Verwendung von Filtern gegen unerwünschte *EMail*-Nachrichten Gedanken zu deren linguistischen Grundlagen.

Die Informationsextraktion beschäftigt sich mit Ansätzen, die spezifische Informationen wie Objekte, Beziehungen oder Ereignisse in natürlichen Texten identifizieren und in schematische Repräsentationen umwandeln. Insbesondere die während der *Message Understanding Conference*[12] gestellten Aufgaben haben diesen Aspekt der Computerlinguistik in den letzten zwanzig Jahren stark vorangetrieben [63]. Mittlerweile erreichen die besten Systeme auch in komplexeren Sätzen Erfolgsraten von über 90 % Präzision und Vollständigkeit[13] bei der Identifikation einfacher Fakten wie Umsatzzahlen oder Unternehmensnamen.

Mit dem Auffinden von Dokumenten in Textsammlungen beschäftigt sich das oft mit ‚Informationsbeschaffung' übersetzte *information retrieval*. Dabei handelt es sich um eine der ältesten Anwendungen der Computerlinguistik, die eine Suche nach relevanten Daten in großen Text-

[12] Eine Serie von Konferenzen zum kompetitiven Vergleich von Methoden der Informationsextraktion.

[13] Präzision bezeichnet den Anteil richtiger Ergebnisse im Verhältnis zu allen erhaltenen Ergebnissen. Vollständigkeit steht für den Anteil der erhaltenen richtigen Ergebnisse im Verhältnis zu allen richtigen Ergebnissen (siehe Abschnitt 2.3.4 auf Seite 32).

sammlungen, die Gruppierung von Texten zu Themengebieten sowie die Identifikation von Schlüsselwörtern und die automatische Erstellung von Zusammenfassungen ermöglicht [130]. Dazu werden Suchanfragen und Textkorpora in linguistische Repräsentationsformen umgewandelt und über verschiedene Kriterien wie der Häufigkeit von Worten oder den semantischen Kontext miteinander verglichen.

Der populär gewordene Begriff des *Text Mining* führt die Informationsbeschaffung und -extraktion mit weiteren Bereiche der Computerlinguistik wie der Textklassifikation [41] zusammen. Einer Vielzahl von Definitionen des *Text Mining*-Begriffs folgend handelt es sich nicht einfach um eine Erweiterung der genannten Verfahren. Vielmehr sollen mittels linguistischer Methoden Fakten gesammelt und durch Datenintegration, Auswertung und Visualisierung zueinander in Bezug gesetzt werden [89, 161]. Ziel ist es, dem Benutzer neue Zusammenhänge oder Hypothesen aus vorhandenem Wissen aufzuweisen.

Automatische Entwicklung von Hypothesen

Einige der ersten erfolgreichen Anwendungen der literaturbasierten Generierung von Hypothesen stammt aus dem biomedizinischen Feld. Seit mehr als 15 Jahren argumentiert Don Swanson, dass die Kombination von existierendem, aber nicht miteinander assoziierten Literaturwissen zu neuen Erkenntnisse führen kann [148]. Es kommt vor, dass eine Publikation die Fakten *A* und *B* miteinander in Verbindung bringt, eine andere wiederum *B* mit *C* assoziiert. Falls ein Zusammenhang zwischen *A* und *C* bisher in der Literatur noch nicht dokumentiert ist, kann die Aufdeckung dieser versteckten Verbindung zu neuem Wissen führen.

Swansons erste Arbeiten basierten auf einer zufälligen Entdeckung während des Studiums der Literatur zur Raynaud-Krankheit, einer Durchblutungsstörung der Extremitäten, und den Wirkungen von Fischtran. Dabei stellte er verschiedene Eigenschaften *B* bei Raynaud-Patienten (der Krankheit *A*) fest: zum Beispiel eine hohe Viskosität des Blutes sowie eine leichte Aggregation der Blutplättchen. Hingegen verringert Fischöl (der Wirkstoff *C*) durch Eicosapentaensäure den Blutdruck und vermindert die Aggregation von Blutplättchen. Zudem erweitert Fischöl die Blutgefäße, welche durch die Raynaud-Krankheit verändert werden. Die Verbindungen *AB* und *BC* waren bereits in der Literatur vorhanden, nur hatte noch

niemand die Hypothese *AC* aufgestellt und überprüft [148]. Angesichts der Menge an verfügbarer Literatur und der begrenzten Zeit, die einem Experten zur Verfügung steht, um sich alleine in seinem Fachgebiet ausreichend zu informieren, ist die Existenz solcher fehlender Verbindungen leicht verständlich. Wie auch seine spätere Entdeckung eines Zusammenhangs zwischen einigen Formen der Migräne und einer Magnesiumdefizienz sind diese Ergebnisse mittlerweile experimentell verifiziert [149]. Nachdem Swanson zeigen konnte, dass Entdeckungen durch die Verbindung von vorher nicht assoziierten Strukturen, Domänen oder Wissensgebiete auch literaturbasiert erfolgen können, fanden computerlinguistische Methoden zunehmend Verwendung[14]. Das in der Folge der Entdeckungen entwickelte Programm *Arrowsmith* zur Vereinfachung der bibliographischen Analyse von Wortfrequenzen in Titeln und Kurzzusammenfassungen war nur teilweise automatisiert: eine ungefilterte Suche konnte schnell zu einer kombinatorischen Explosion der Begriffe führen und benötigte daher eine ständige Kontrolle durch den Benutzer. Mittlerweile gibt es aber verschiedene Methoden, das Verfahren durch linguistische Filter – wie die Überprüfung der semantischen Beziehung gefundener Konzepte – weitgehend zu vereinfachen und zu automatisieren [99, 116, 165].

Keine der Methoden zur selbstständigen Erstellung von Hypothesen ist als das perfekte Mittel zur Generierung neuen Wissens gedacht. Ihr Ziel ist die Unterstützung von Experten, nicht deren Ersetzung, da fast alle Programme manuelle Eingriffe bei der Erstellung der ersten Anfragen, bei der Auswahl der sinnvollen Parameter und vor allem bei der Auswertung und Evaluierung der vorgeschlagenen Hypothesen benötigen. Es geht nach Valdés-Pérez bei der literaturgestützten Hypothesengenerierung nicht um die automatisierte Suche von Mustern in Daten, sondern um die Entdeckungen in der Wissenschaft, oder…

> — *„…the generation of novel, interesting, plausible and intelligible knowledge about the objects of study. "[157]*

[14] Eine Übersicht von computergestützten Methoden zur Entdeckung neuen Wissens allgemein findet sich in Langley [92] und Valdés-Pérez [157], ein Schwerpunkt zur literaturbasierten Hypothesengenerierung in Weeber [160].

Bioinformatische Anwendungsmöglichkeiten der Computerlinguistik

In der Bioinformatik gilt die Computerlinguistik nicht nur als Hilfsmittel für andere Programme, sondern als eigenständiges Forschungsgebiet. Linguistische Methoden finden seit mehr als zwanzig Jahren ihre Anwendung in der Biologie, insbesondere zur formellen Repräsentation biologischer Sequenzen oder der Analyse dreidimensionaler Strukturen von RNA-Molekülen [137]. Die Computerlinguistik als Methode zur Verarbeitung von Sprache hat sich in den letzten zehn Jahren nach und nach in der Bioinformatik durchgesetzt[15]. Mittlerweile fand mit dem *KDD Challenge Cup*[16] analog zur *Message Understanding Conference* auch ein Wettbewerb zum Vergleich von Methoden der Informationsextraktion aus biologischen Publikationen statt [170].

Bei der Informationsbeschaffung und den damit verwandten Methoden reichen die Anwendungen von der Suche in großen Literaturdatenbanken wie *PubMed* [169] über die Wissensrepräsentation [12] bis zur automatischen Annotation neuer Sequenzen basierend auf der Beschreibung verwandter Sequenzen [3, 4, 51, 120]. Andere Anwendungen umfassen die Erweiterung von Ontologien [18, 20] oder beschäftigen sich mit der Verbesserung bestehender Algorithmen wie PSI-BLAST unter Einbeziehung der Ähnlichkeit von Sequenzannotationen [36].

Die Anzahl an Publikation zur Informationsextraktion ist fast unüberschaubar und spiegelt den Bedarf wieder, biologischen Objekten und Prozessen automatisch Eigenschaften zuordnen zu können. Einige Beispiele dafür sind die Vorhersage der subzellulären Lokalisation von Proteinen [144], der Identifikation von Inhibitoren biologischer Vorgänge [117], der Sammlung von Informationen zu Proteinstrukturen [46] und die Extraktion von Protein- oder Gennamen [90]

Besonders intensiv untersucht sind die komplexeren Aufgaben der Extraktion von Protein-Protein- und Gen-Gen-Interaktionen [16, 53, 80, 150, 152]. Während sich die ersten Auswertungen von Interaktionen auf kleine Teilgebiete beschränkten, ermöglichte das *PubGene*-Programm erstmals die Analyse der gemeinsamen Nennung von Gennamen in allen Titeln und Kurzzusammenfassungen der *PubMed*-Datenbank [145]. Auf-

[15] Eine ausführliche Übersicht der verschiedenen Themengebiete und der verwendeten Methoden findet sich in de Bruijn und Martin [25] und Hirschman *et al.* [71].

[16] Im Rahmen der Konferenz für *Knowledge Discovery and Data Mining* (KDD).

wändige Verfahren wie *Empathie* nutzen komplexe Regelsysteme zur Extraktion von Enzyminteraktionen und dem Aufbau metabolischer Netzwerke [73], oder erstellen eine kontext-freie Grammatik zur Identifikation von Protein-Protein-Interaktionen [151]. Im Vergleich zu den im Bereich klassischer Informationsextraktionsaufgaben publizierten Ergebnissen von über 90 % Präzision und Vollständigkeit erscheinen die erreichten Erfolgsraten von circa 80 % Präzision bei 50 % Vollständigkeit eher gering. Allerdings sind diese Aufgaben wesentlich komplexer, da jeweils zwei Entitäten *und* deren Beziehung zueinander extrahiert werden müssen. Dennoch reichen Präzision und Vollständigkeit oft schon aus, um zum Beispiel die aus *Yeast Two Hybrid*-Experimenten [60] gewonnen Daten zu ergänzen.

Die beschriebenen Verfahren reichen von einfachen Systemen zur Ermittlung von Wortfrequenzen und der Mustererkennung über komplexere Methoden, die Wortbezüge analysieren, bis hin zu Anwendungen, die auch Zusammenhänge über Satzgrenzen zu verfolgen versuchen. Alle Anwendungen spezialisieren sich jedoch auf ein Themengebiet und eine bestimmte linguistische Methode. Nach wie vor fehlt ein einfaches, modulares System, welches sich ohne großen Aufwand für eine Vielzahl unterschiedlicher Domänen verwenden lässt und sich dabei gleichzeitig den Anforderungen eines Benutzers anpasst.

1.1 Aufgabenstellung

Ziel dieser Dissertation war die Entwicklung eines Programms zur automatischen Erweiterung der BRENDA-Datenbank. Als Prototyp sollte das Verfahren Informationen zu Krankheiten, die kausal mit Enzymklassen in Verbindung stehen, mit hoher Präzision aus der wissenschaftlichen Literatur extrahieren. Das dazu notwendige Verfahren sollte folgende Schritte enthalten:

- Die Erstellung einer geeigneten Sammlung von zu untersuchenden Dokumenten, dem sogenannten Textkorpus.
- Die Erkennung der in der BRENDA-Datenbank vorhandenen Enzymnamen im Textkorpus.
- Die Auswahl eines medizinischen Vokabulars und die Zuordnung dieser medizinischen Begriffe zu einzelnen Dokumenten oder Bausteinen der Dokumente.
- Die Entwicklung einer zuverlässigen Zuordnung zwischen identifizierten Enzymnennungen und damit kausal zusammenhängenden Krankheitsbegriffen mit einem Schwerpunkt auf der Genauigkeit der Zuordnung.
- Eine Analyse der Präzision und Vollständigkeit der Zuordnungen sowie ein Vergleich mit in anderen Datenbanken enthaltenen Annotationen.

Zwei Schwerpunkte bei der Entwicklung des Verfahrens stellten die Anpassungsmöglichkeit auf andere enzymrelevante Informationen sowie die Entwicklung von Methoden zur Vereinfachung einer manuellen Auswertung der extrahierten Daten dar. Zusätzlich sollten Methoden zur automatischen Generierung von Hypothesen entwickelt werden, um so interessanten Zusammenhängen zwischen Krankheiten und Enzymklassen nachgehen zu können.

2 THEORIE & METHODEN

> This one's tricky. You have to use imaginary numbers, like eleventeen.
>
> *aus: Calvin & Hobbes*

Die in dieser Arbeit entwickelte Methode zur automatischen Extraktion krankheitsrelevanter Enzyminformationen aus wissenschaftlichen Publikationen sowie die Präsentation und Auswertung der Daten lässt sich in mehrere Abschnitte gliedern:

- *Aufbau und Analyse eines Textkorpus:* In Abschnitt 2.1 wird der Aufbau einer geeigneten Sammlung von auszuwertenden Dokumenten, ihre lexikalische Analyse und die Identifikation von Konzepten beschrieben.
- *Automatische Erkennung von Enzymnamen:* In Abschnitt 2.2 findet sich die Identifikation von Enzymnamen in den vorliegenden Texten unter Verwendung eines aus der BRENDA-Datenbank erstellten Namenslexikons.
- *Zuordnung von Krankheitsbegriffen zu Enzymklassen:* Abschnitt 2.3 beschreibt zwei verschiedene Ansätze zur automatischen Zuordnung von Enzymklassen zu Krankheiten aufgrund einer gemeinsamen Nennung in Publikationen. Ein Übersicht der Kriterien zum Qualitätsvergleich der Zuordnungen schließt diesen Teil ab.

- *Visualisierung und Auswertung der Netzwerke:* Die Verfahren zur Analyse der Netzwerke aus Krankheiten und Enzymklassen sowie deren Darstellung sind in Abschnitt 2.4 dargestellt.

Das Diagramm 2.1 auf der nächsten Seite skizziert den Arbeitsablauf und verweist auf die entsprechenden Abschnitte im Methodenteil.

2.1 Aufbau & Analyse eines Textkorpus

2.1.1 Erstellung eines Textkorpus

Aus der *PubMed*-Datenbank abgerufene Kurzzusammenfassungen bildeten die Grundlage zum Aufbau des in dieser Arbeit verwendeten Textkorpus. Die Abfrage erfolgte mittels automatisch generierter Suchanfragen. Um die Anzahl der Artikel ohne Krankheits- und Enzyminformationen möglichst gering zu halten, bestand die jeweils erste Anfrage nur aus dem empfohlenen Enzymnamen und der dazugehörigen *Enzyme Commission* (EC)-Nummer (siehe Abschnitt 2.2.1 auf Seite 24) in Kombination mit einem der krankheitsbezogenen *Medical Subject Headings* (MESH). Bei diesen MESH-Begriffen handelt es sich um ein kontrolliertes, hierarchisch organisiertes Vokabular von 22 000 Schlagwörtern aus biomedizinischen, klinischen und experimentellen Bereichen, mit denen die Kuratoren der *PubMed*-Datenbank archivierte Publikationen annotieren. Auf Krankheiten beziehen sich dabei die Kategorien von C01 (bakterielle Infektionen) bis C21 (umweltbedingte Erkrankungen) mit insgesamt 3 747 Begriffen. Die Entwicklung des Prototypen sollte sich auf Erkrankungen des Menschen beschränken, daher fanden die Kategorien C22 (Erkrankungen von Tieren) sowie C23 (Pathologische Symptome) keine Verwendung, obwohl sie ebenfalls der Kategorie ‚Krankheiten' untergeordnet sind. Alle benutzten Kategorien und ihre Bezeichnung sind im Anhang A.2 auf Seite 126 aufgelistet.

Einträge ohne vollständige Kurzzusammenfassung in englischer Sprache und dem MESH-Schlagwort *human* für Humanstudien wurden nicht berücksichtigt (siehe Abbildung 2.2 für ein Abfragebeispiel). Fanden sich mehr als 1 000 übereinstimmende Publikationen je Enzymklasse, so erwies sich eine schrittweise Einschränkung des Publikationszeitraums als

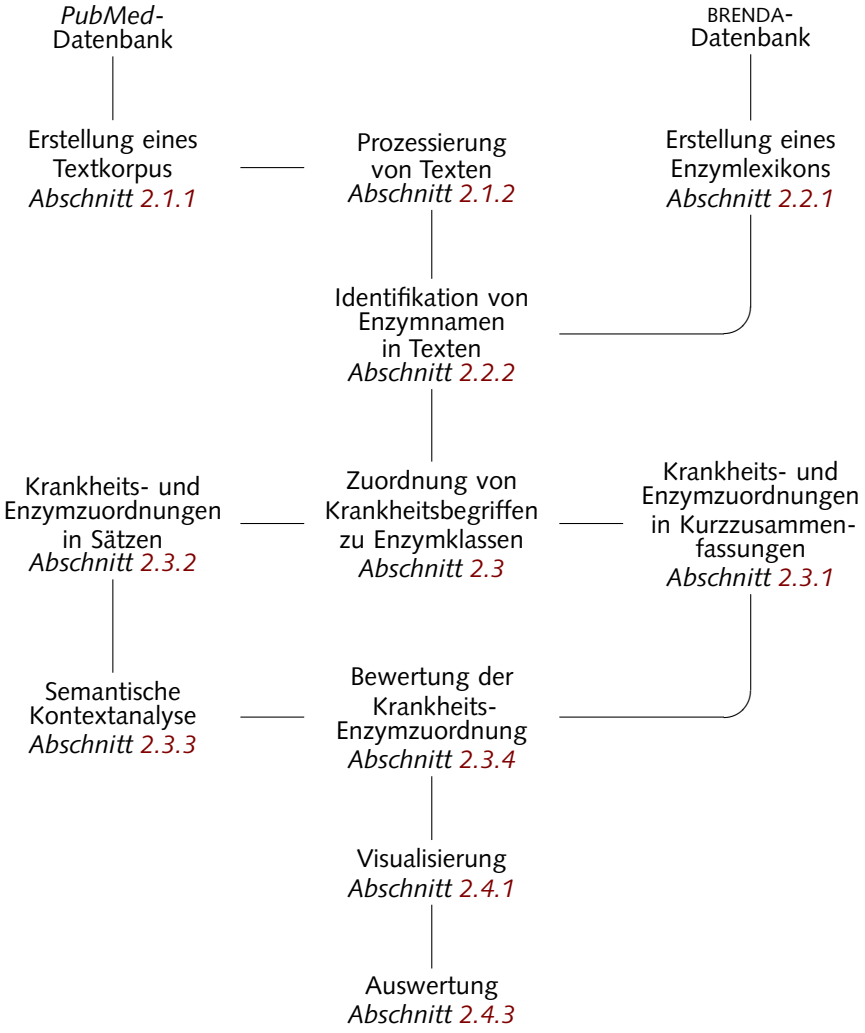


Abbildung 2.1 · Übersicht der Programmabläufe für die Gewinnung eines Textkorpus, die Analyse der Dokumente und die Zuordnung von Enzymen zu Krankheiten und deren Auswertung. Verweise auf die entsprechenden Abschnitte sind *kursiv* geschrieben.

sinnvoll, um eine gleichmäßige Verteilung von Dokumenten zu den verschiedenen Enzymklassen zu gewährleisten. Wurden weniger als zehn Dokumente gefunden, erfolgte eine erneute Anfrage unter Einbeziehung der synonymen Enzymnamen. Alle derart gesammelten Dokumente wurden für die weitere Analyse in einer Datenbank lokal gespeichert. Im weiteren Verlauf wird diese Sammlung von *PubMed*-Dokumenten als Textkorpus bezeichnet.

2.1.2 Prozessierung von Texten

Die lexikalische Analyse der *PubMed*-Dokumente stellte den ersten Prozessierungsschritt des Textkorpus dar. Verwendung fanden hierbei der Titel, die Kurzzusammenfassung und die Publikation beschreibende MESH-Schlagwörter. Nach der Entfernung redundanter MESH-Begriffe wurde der Inhalt der drei Bereiche zusammen mit allen Zwischenstufen der Analysen in einer Datenbank gespeichert.

Trennung in Sätze und Bildung von Token

Der erste Bestandteil der lexikalischen Analyse stellte die Zerlegung der Texte in einzelne Bausteine, sogenannte *Token*, über die Bestimmung von Wort- und Zahlengrenzen dar. Zur Generierung von *Token* wurden der bearbeitete Text an jeder Punktuation mit vorhergehendem Kleinbuchstaben und nachfolgendem Leerzeichen voneinander getrennt. Durch die Zusammenlegung von Abschnitten mit weniger als 15 Zeichen mit dem jeweils nachfolgenden Satz konnte eine unerwünschte Satztrennung nach mit Punkten versehenen Abkürzungen vermieden werden.

```
hasabstract AND human[MESH] AND English[lang]
AND (1.1.1.1 [EC] OR "alcohol dehydrogenase"
OR "nadh-alcohol dehydrogenase" OR ...)
AND "musculoskeletal diseases" [MESH]
```

Abbildung 2.2 · Beispiel einer automatischen Abfrage der *PubMed*-Datenbank für die EC-Nummer 1.1.1.1, (Alkohol Dehydrogenase), in Kombination mit dem MESH-Begriff *musculoskeletal diseases*.

Nach der Aufteilung in Sätze erfolgte die Entfernung der in Tabelle 2.1 angegebenen Punktuation, Sonderzeichen, Klammerungen und beidseitig von Leerzeichen begrenzte Bindestriche. Nach einer Umwandlung der Großbuchstaben in Kleinbuchstaben konnte jede von Leerzeichen begrenzte Abfolge von Nicht-Leerzeichen als *Token* definiert werden[1]. Abbildung 2.3 zeigt einen Auszug aus einer Kurzzusammenfassung vor und nach der Umwandlung in *Token*. Auf dieser Repräsentation der *PubMed*-Dokumente beruhte die in Abschnitt 2.2.2 beschriebene Identifikation von Enzymnamen.

Wörter und Stammformen

Nach der Entfernung numerischer Ausdrücke definierten die übrigen alphanumerischen *Token* das Vokabular des Textkorpus. Um dieses Vokabular auf für den Vergleich von Texten und die Erstellung von Kurzzusammenfassungen (siehe Abschnitt 2.4.2) relevante Einträge zu reduzieren, wurden die Wörter durch eine Stoppwortliste gefiltert. Dabei handelt es sich um eine Sammlung von etwa 300 Wörtern mit niedrigem Informationsgehalt, die zu einer Charakterisierung der Texte nicht notwendig sind und bei einem Vergleich von Texten nur eine geringe Rolle spielen[2]. Die von der *PubMed*-Datenbank bereitgestellte statische Stoppwortliste (siehe Abschnitt A.1 im Anhang) wurde zusätzlich dynamisch um Wörter ergänzt, die in mehr als 25 % der Texte im Textkorpus vorkamen. Als

[1] Dabei kann es vorkommen, dass zusammengehörige Wörter in einzelne Komponenten zerteilt werden, z. b. HER2/HER3 *heterodimers* → HER2, HER3, *heterodimers*.

[2] Typische Stoppwörter in der deutschen Sprache sind zum Beispiel ‚und', ‚oder' sowie die Artikel ‚der/die/das'.

Tabelle 2.1 · Während der Textprozessierung gesondert behandelte Zeichen. Punktuationen fanden Verwendung bei der Unterteilung der Kurzzusammenfassungen in einzelne Sätze; Sonderzeichen und Klammerungen beschreiben die in den Texten durch Leerzeichen ersetzten Symbole.

Regel	Zeichen
Punktuation	. : ; ! ?
Sonderzeichen	$ * % < > -/
Klammerung	[] () {}

— „*The human CuZn superoxide dismutase (superoxide dismutase 1), a key enzyme in the metabolism of oxygen free-radicals, is encoded by a gene located on chromosome 21 in the region 21 q 22.1 – known to be involved in Down's syndrome. A gene dosage effect for this enzyme. . . "*

Satz	*Token*-Liste
1	the human cuzn superoxide dismutase superoxide dismutase 1 a key enzyme in the metabolism of oxygen free-radicals is encoded by a gene located on chromosome 21 in the region 21 q 22.1 known to be involved in down's syndrome
2	a gene dosage effect for this enzyme. . .

Abbildung 2.3 · Beispiel für die Aufteilung eines Textes in Sätze und die Umwandlung in eine Liste von *Token*. Jede von Leerzeichen begrenzte Folge von Zeichen stellt ein *Token* dar.

letztes erfolgte eine Filterung von Wörtern, die im gesamten Textkorpus nur einmal vorkamen und dadurch bei einem Textvergleich nicht hilfreich sind.

Eine weitere Reduktion und Vereinheitlichung des Vokabulars konnte mit der Anwendung des *Porter Stemmer*-Algorithmus erreicht werden, welcher verschiedene Flexionen eines Wortes auf die jeweilige Stammform zurückführt [115]. Jedes Wort wurde mit seiner Position und der dazugehörigen Stammform der Datenbank hinzugefügt. Ein Beispielsatz und seine Repräsentation durch Wörter und Stammformen findet sich in Tabelle 2.3 auf Seite 23. Die erhaltenen Wörter und Stammformen wurden nur für den Textvergleich und die Erstellung der Kurzzusammenfassungen benötigt (siehe Abschnitt 2.4.2 und 2.4.2 auf Seite 36).

Konzepte

Die Identifikation von Konzepten im Textkorpus ermöglichte eine Vereinheitlichung des verwendeten Vokabulars. Unterschiedliche Wörter, Wortfolgen oder Stammformen können die gleiche Bedeutung haben, was eine Repräsentation solcher Synonyme durch ein einheitliches Konzept notwendig macht. Das Resultat einer Repräsentation durch Konzepte ist die

Umwandlung der freien Sprache der Publikationen in ein kontrolliertes Vokabular. Für eine Identifikation von Konzepten in den Kurzzusammenfassungen wurde die *MetaMap*-Software verwendet, ein Programmpaket der *Semantic Knowledge Representation*-Gruppe [10]. *MetaMap* analysiert einzelne Sätze mit Hilfe linguistischer Methoden und identifiziert darin Konzepte der UMLS-Ontologie [6, 8]. In der verwendeten 13. Auflage der Ontologie enthält über 1,5 Millionen Bezeichnungen aus mehr als 60 unterschiedlichen Vokabularen, die zu 775 000 Konzepten gruppiert vorliegen. Zusätzlich ist jedes Konzept einem oder mehreren semantischen Feldern zugeordnet. Die semantischen Felder – eine Übersicht findet sich in Anhang A.3 auf Seite 127 – beschreiben die Eigenschaften oder Bedeutung eines Konzepts. So ist zum Beispiel der Begriff ‚Ratte' dem Feld ‚Säugetier' zugeordnet, eine ‚Polymerasekettenreaktion' dem Feld ‚molekularbiologische Labortechnik'. Der beispielhaft in Tabelle 2.2 gezeigte Ablauf für die Konzeptzuweisung besteht aus:

- einer Analyse des Satzes mittels des *Xerox Part-of-Speech-Taggers* [44] der die Wortart einzelner Wörter bestimmt (also Substantive, Adjektive, Verben und andere Bestandteile der Syntax identifiziert).
- der Annotation von Substantivphrasen aus zusammengehörenden Wörtern, z. B. der Wortfolge ‚chronische Entzündung der Lunge'.
- einer Generierung von Varianten der Substantivphrasen durch Bildung von lexikalischen Abwandlungen (Abkürzungen, Synonyme, Derivationen und Flexionen[3]) und Permutation der Wortreihenfolge.
- einer Zuordnung der Varianten zu Konzepten der UMLS-Ontologie und der Berechnung einer Bewertungszahl, welche die Konfidenz in die Zuordnung wiedergibt. Dieser Wert ist abhängig vom Grad der Übereinstimmung zwischen Variante und Konzept sowie der Anzahl an notwendigen Modifikationen der ursprünglichen Substantivphrase [7].

[3] Bsp: *Augen* sind eine Flexion von *Auge* mit dem Synonym *Oculus*, eine Ableitung (Derivation) davon ist *ocular*.

Tabelle 2.2 · Beispiel für den Prozessierungsablauf von Sätzen zu Konzepten durch das *MetaMap*-Programm. Nach einer Syntaxanalyse des Textes werden zusammenhängende Phrasen markiert. Die Generierung von Varianten führt zu einer Liste von Kandidaten, wobei die besten Übereinstimmungen mit Konzepten der UMLS-Ontologie ausgegeben werden. Die Abkürzung NOS steht hier für *not otherwise specified*.

Satz	*the*	*local*	*anesthetic*	*bupivacaine*	*is*	*cardiotoxic*
Syntax		Adjektiv	Adjektiv	Substantiv	Verb	Adjektiv
Phrase	the local anesthetic bupivacaine				is	cardiotoxic
Variante		local anesthethics local anaesthethic local anesthethic local anaesthethics local	anaesthetic anesthetist anaesthetist anesthetized anaesthetised anesthetists anesthetic anesthetics anaesthetics	bupivacaine		cardiotoxic cardiotoxicity
Kandidat		Local anaesthetic Local anaesthethic/NOS Local Anaesthethic/NOS Anesthetist		Bupivacaine		Cardiotoxicity
Konzept		Local anaesthetic/NOS		Bupivacaine		Cardiotoxicity

Die Standardparameter für *MetaMap* legen einen Schwerpunkt auf hohe Genauigkeit. Um die Rate an falsch-positiven Zuordnungen zu verringern, war eine Einschränkung auf die Verwendung eindeutiger Akronyme und Synonyme notwendig (*MetaMap*-Option *-u*). Variationen in der Wortreihenfolge waren erlaubt (Option *-i*), Lücken in den Konzeptzuordnungen nicht: dadurch war eine unvollständige Abdeckung einer Phrase mit Konzepten nur zu Beginn und am Ende der Phrase möglich. So wäre für *acute ear infection* die Zuordnung des Konzepts *ear infection* erlaubt, die von *acute infection* dagegen nicht. Die Verwendung des strikten Modells während der Präprozessierung der UMLS-Ontologie (Option *-A*) führte zur Entfernung von Konzepten, die zu generisch, zu lang oder zu komplex sind, oder für die aufgrund ihrer internen Struktur eine Identifikation nicht zu erwarten ist [9]. Alle Zuordnungen wurden zusammen mit der Bewertung und ihrer Position im Text in der Datenbank gespeichert. Ein Beispielsatz zusammen mit den verschiedenen Repräsentationen durch Wörter, Stammformen und Konzepte findet sich in Tabelle 2.3 auf der nächsten Seite.

Repräsentation der Dokumente als Vektor

Um einen Vergleich von Texten durchführen zu können, wurde nach dem von Salton entwickelten *Vector Space Model* [131] eine Vektorrepräsentation der Dokumente gewählt. Das Verfahren beruht auf einer Darstellung von Dokumenten über ihre Terme. Als Term können sowohl Wörter, Stammformen oder auch Konzepte verwendet werden.

Jedes Dokument j wird durch einen Vektor D dargestellt, wobei eine Komponente i des Vektors das Gewicht eines Terms in diesem Dokument darstellt. Das Gewicht repräsentiert die Bedeutung des Terms. Die Länge eines einzelnen Dokumentenvektors entspricht dabei der Größe des vorliegenden Lexikons. Alle Dokumentenvektoren zusammen spannen einen k-dimensionalen Raum auf, den Vektorraum. Bei diesem Ansatz werden grammatikalische Strukturen vernachlässigt, das Gewicht eines Terms ist unabhängig von dessen Position innerhalb des Textes.

Eine Methode zur Berechnung von Termgewichten stellen die Modelle des SMART-Systems dar, bei denen das Gewicht v_{ij} des Terms i in Dokument j in drei Schritten berechnet wird [37, 128, 129].

Tabelle 2.3 · Beispiel für die Umwandlung von Texten in Wörter, Stammformen und Konzepte.

> *— „The cells were 52 % positive for peroxidase staining and manifested strongly positive activity of alpha-naphthyl acetate esterase, which was completely inhibited by sodium fluoride."*

Wörter	Stammformen	Konzepte
cells	cell	the cell
positive	posit	positive
peroxidase	peroxidas	peroxidase
staining	stain	staining
manifested	manifest	*not found*
positive	posit	positive
activity	activ	activity
alpha-naphthyl	alpha-naphthyl	acetate esterase, alpha-naphthyl
acetate	acet	
esterase	esteras	
completely	complet	*not found*
inhibited	inhibit	inhibition
sodium	sodium	fluoride, sodium
fluorid	fluorid	

Die hier verwendete Variante entspricht dem ‚atc'-Modell des SMART-Systems:

$$v_{ij} = TF_{ij} \cdot IDF_i \cdot N_j$$

Dabei steht TF für die erweiterte Termfrequenz, welche die relative Bedeutung des Terms i in Dokument j wiedergibt:

$$TF_{ij} = \begin{cases} 0.5 + 0.5 \cdot \frac{f_{ij}}{max_i(f_{ij})}, & f_{ij} > 0 \\ 0, & f_{ij} = 0 \end{cases}$$

mit f_{ij} für die Häufigkeit von Term i in Dokument j, und $max_i(f_{ij})$ für die höchste Häufigkeit eines Terms dieses Dokuments. Dadurch erfolgt eine Normierung der unterschiedlichen Länge der betrachteten Dokumente.

Die inverse Dokumentenfrequenz IDF, welche die Bedeutung des Terms i im Textkorpus allgemein angibt, berechnet sich wie folgt:

$$IDF_i = \log\left(\frac{d}{d_i}\right)$$

mit d für die Gesamtzahl an Dokumenten im Textkorpus und d_i für die Anzahl von Dokumenten, in denen Term i vorkommt. Als letztes wird eine Kosinusnormalisierung durchgeführt:

$$N_j = \frac{1}{\sqrt{\sum_{i=1}^{m}(IDF_i \cdot TF_{ij})^2}}$$

wobei m für die Anzahl unterschiedlicher Terme in Dokument j steht. Das so bestimmte Gewicht bildet eine Komponente des Dokumentenvektors $D_j = v_{1j}, v_{2j}, \ldots, v_{nj}$.

2.2 Automatische Erkennung von Enzymnamen

2.2.1 Erstellung eines Enzymlexikons

Vor dem Aufbau des Textkorpus wurde ein Lexikon von Enzymnamen erstellt, um diese für *PubMed*-Abfragen und die Namensidentifikation in den abgerufenen Texten verwenden zu können. Das Namenslexikon enthielt für jede EC-Nummer in der BRENDA-Datenbank (Stand Juli 2002) den empfohlenen Enzymnamen und alle Synonyme mit mehr als vier Buchstaben. Das System der Enzymklassen basiert auf der Nomenklatur der Enzymkommission der *International Union of Biochemistry and Molecular Biology* (IUBMB), welche die unterschiedlichen Enzyme entsprechend der katalysierten Reaktion und ihrer Substratspezifität benennt [153]. Die EC-Nummer besteht aus vier durch Punktierung voneinander getrennten Zahlen, wobei die ersten drei eine zunehmend detailliertere Beschreibung der Enzymfunktion angeben; die erste Ziffer gibt die Hauptklasse wieder. Die letzte Zahl entspricht einer laufenden Nummer. Bei den Synonymen handelt es sich um eine Sammlung von in der Literatur verwendeten Namen für Enzyme dieser Enzymklasse. Der empfohlene Name geht oft auf eine verkürzte Form des systematischen Namens eines Enzyms zurück oder entspricht dem gebräuchlichsten Synonym.

Alle Enzymbezeichnungen wurden analog zur Generierung von *Token* in Abschnitt 2.1.2 auf Seite 17 verarbeitet, um sie später mit den Texten vergleichen zu können: Sonderzeichen, Klammerungen, Punktuation und von Leerzeichen begrenzte Bindestriche wurden entfernt und alle Großbuchstaben in Kleinbuchstaben umgewandelt

2.2.2 Identifikation von Enzymnamen in Texten

Mittels des Enzymlexikons erfolgte die automatische Durchsuchung des gespeicherten Textkorpus nach Nennungen von Enzymen. Für jedes Dokument wurden die darin gefundenen EC-Nummern, empfohlenen Enzymnamen und Synonyme gespeichert, wobei eine Auflösung von Ambiguitäten über das Prinzip der größten Übereinstimmung stattfand – bei mehreren Übereinstimmungen zwischen *Token* des Dokuments und Enzymnamen innerhalb des gleichen Textintervalls erhielt der Name mit der größeren Abdeckung des Intervalls den Vorrang. Im folgenden Beispiel wird eine im Text gefundene Nennung einer *tyrosine kinase* durch eine Expansion zu *protein tyrosine kinase pp56clk* ersetzt:

...activity of	*protein*	*tyrosine*	*kinase*	*pp56lck*	in isolated...
		tyrosine	kinase		
	protein	tyrosine	kinase		
	protein	tyrosine	kinase	pp56clk	

Eine Auflösung nicht eindeutiger Zuordnungen aufgrund mehrdeutiger Enzymnamen war teilweise durch eine Kontextanalyse möglich. Wenn an anderer Stelle des gleichen Dokuments ein eindeutiger Enzymname der gleichen EC-Nummer identifiziert werden konnte, erfolgte eine Annotation nur mit dieser Enzymklasse.

Ähnlichkeit von Enzymnamen

Um die Möglichkeit zu überprüfen, von den im Enzymlexikon vorkommenden Namen leicht abweichende Schreibweisen in Texten identifizieren und zusammenführen zu können, wurden die Levenshtein-Distanz aller Einträge des Lexikons zueinander berechnet. Die Levenshtein-Distanz entspricht der Anzahl von Insertionen, Deletionen und Substitutionen, die

notwendig sind, um eine Anfangszeichenfolge s in eine Zielzeichenfolge t umzuwandeln – je größer die Distanz, desto unterschiedlicher die Zeichenfolgen [95]. Dazu wurden nach der Umwandlung der Enzymnamen in *Token* zusätzlich alle arabischen und römischen Ziffern in Worten ausgeschrieben. Nach Entfernung der Zeichen ‚+' und ‚–' erfolgte als letzter Schritt der Austausch aller weiteren nicht alphabetischen Zeichen durch ein einheitliches Symbol ('#').

2.3 Zuordnung von Krankheitsbegriffen zu Enzymklassen

Nach der lexikalischen Analyse des Textkorpus und der Identifikation von Enzymnennungen in den einzelnen Dokumenten konnten im nächsten Schritt Zusammenhänge zwischen Krankheitsbegriffen und Enzymklassen hergestellt werden. Die Zuordnung beruhte auf der Auswertung von Kollokationen der Krankheitsbegriffe und Enzymnamen innerhalb eines Textabschnitts. Als Kollokation bezeichnet die Linguistik spezifische und charakteristische Kombinationen von Wörtern, die semantisch zueinander passen – wie zum Beispiel der ‚geneigte Leser'. In der statistischen Definition entfällt diese Beschränkung, jede häufige Kombination von Wörtern innerhalb eines Textabschnitts wird als Kollokation bezeichnet. In dieser Arbeit wurden dazu gemeinsame Nennungen in Kurzzusammenfassungen (Abschnitt 2.3.1) oder Einzelsätzen (Abschnitt 2.3.2) ausgewertet.

2.3.1 Krankheits- und Enzymzuordnungen in Kurzzusammenfassungen

Die erste Art der Zuordnung basierte auf dem gemeinsamen Auftreten eines relevanten MESH-Begriffes mit einem Namen aus dem erstellten Enyzmlexikon (siehe Abschnitt 2.2.1) in einem *PubMed*-Dokument. Als krankheitsrelevant galten in diesem Zusammenhang MESH-Begriffe der schon in Abschnitt 2.1.1 auf Seite 15 für den Aufbau des Textkorpus verwendeten Kategorien. Da verschiedene Krankheiten unterschiedlich gut untersucht sind, variiert die Häufigkeit der verwendeten MESH-Begriffe bei der Annotation deutlich. Für Volkskrankheiten wie Diabetes findet sich

aufgrund der großen Anzahl an Publikationen eine Kollokation mit fast jeder Enzymklasse, während sich nur eine kleinere Anzahl an Veröffentlichungen mit seltenen Krankheiten beschäftigt. Dies machte eine variable Berechnung der Untergrenze von Kollokationen notwendig, ab der eine Zuordnung erfolgen konnte: Einem MESH-Begriff wurde eine Enzymklasse nur dann zugeordnet, wenn die Enzymklasse im Verhältnis zu anderen Enzymen überdurchschnittlich oft mit dem MESH-Begriff vorkam; oder aber mindestens halb so oft wie die Enzymklasse mit der höchsten Zahl an Kollokationen für diesen MESH-Begriff gefunden werden konnte.

> Für den MESH-Begriff ‚Diabetes Mellitus' finden sich zwölf Dokumente, in denen die Hexokinase identifiziert wurde, sechs weitere Dokumente mit einer Nennung der Glukokinase, sowie drei mit dem Enzym Cytochrom-c Oxidase. Die durchschnittliche Anzahl an gemeinsamen Nennungen für ‚Diabetes Mellitus' liegt damit bei sieben, die Hälfte der maximalen Nennungen bei sechs. Damit erfolgt nur für die Cytochrom-c Oxidase keine Zuordnung zu diesem MESH-Begriff.

Als allgemeine Untergrenze für eine Zuordnung zwischen einer Enzymklasse und einem der MESH-Krankheitsbegriffe waren mindestens drei gemeinsame Nennungen in verschiedenen Publikationen notwendig, um falsche Zuordnungen zu vermeiden.

Zusammenlegung unterschiedlicher Annotationen

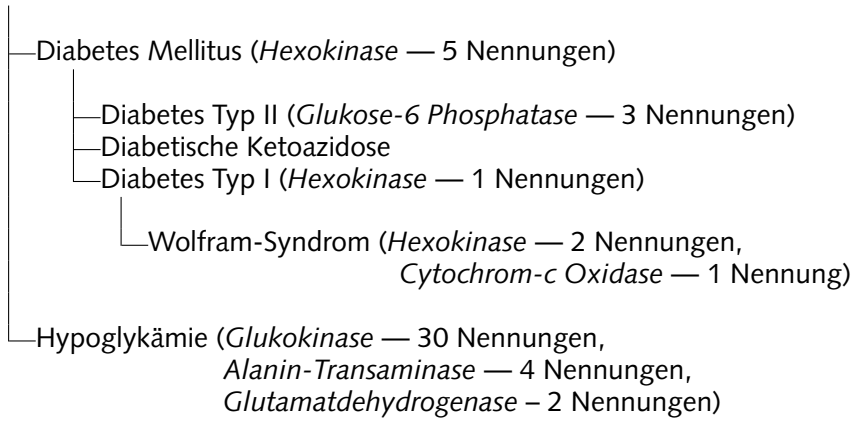
Während die Filterung seltener Kollokationen die Konfidenz einer Zuordnung verbessert, führt dies zu Problemen bei einer ähnlichen, sich nur in ihrer Spezifität unterscheidenden Annotation durch die Experten der *PubMed*-Datenbank. Unterschiedliche Kuratoren vergeben oft andere oder zumindest leicht abweichende Schlagwörter zur Indexierung einer Publikation [2] – während ein Kurator einen Artikel mit dem Schlagwort ‚Augeninfektion' charakterisiert, verwendet ein anderer den spezielleren Begriff ‚bakterielle Augeninfektion'. Dies führt zu unterschiedlichen Zuordnungen mit entsprechend geringerer Häufigkeit. Zur Auflösung solcher Ungenauigkeiten wurde die hierarchische Struktur des MESH-Vokabulars genutzt. Dieses ist in einer Baumstruktur organisiert, bei der

sich allgemeinere Begriffe in der Nähe der Wurzel befinden, spezialisiertere Schlagwörter dagegen in der Nähe der Blätter zu finden sind. Bei der Zuordnung zwischen Enzymklassen und MESH-Begriffen wurde das oben beschriebene Verfahren wie folgt durchgeführt (Abbildung 2.4 auf der nächsten Seite verdeutlicht das Verfahren):

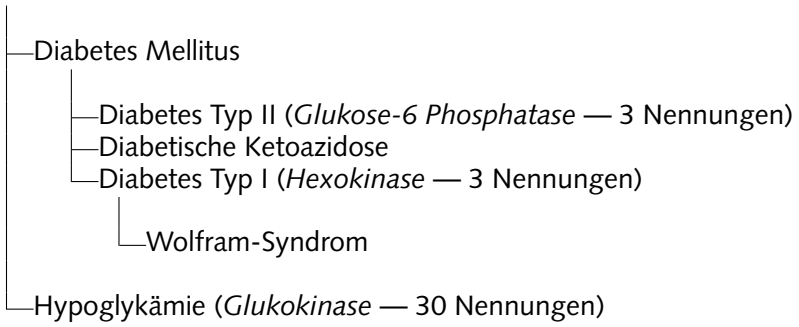
- Die Zuordnung von Enzymklassen zu einem MESH-Begriff erfolgte zuerst für den MESH-Begriff mit der höchsten Spezialisierung.
- Alle Nennungen von Enzymklassen, die dem bearbeiteten MESH-Begriff zugeordnet werden konnten, wurden aus übergeordneten, allgemeineren MESH-Begriffen entfernt. Dadurch fand nur eine Zuordnung zwischen der Enzymklasse und dem am höchsten spezialisierten Krankheitsbegriff statt.
- *Nicht* zugeordnete Enzymklassen – also solche mit einer zu geringen Zahl an Nennungen – wurden zum nächsten, übergeordneten MESH-Begriff verschoben, um somit leicht abweichende Annotationen zusammenzufassen.
- Das Verfahren wurde für den nächsten MESH-Begriff fortgesetzt.

2.3.2 Krankheits- und Enzymzuordnungen in Sätzen

Eine weitere Zuordnung zwischen Krankheitsbegriffen und Enzymklassen fand aufgrund der Kollokation von Krankheiten und Enzymklassen innerhalb eines einzelnen Satzes statt. Die Auswertung eines im Vergleich zu Kurzzusammenfassungen kleinen Textabschnitts wie eines Satzes führt zu einer höheren Genauigkeit der erstellten Zuordnungen [50]. Die Annotation der Kurzzusammenfassungen mit MESH-Begriffen enthielt keine Informationen darüber, welcher Satz der Kurzzusammenfassung zu der jeweiligen Annotation führte. Daher musste für die Identifikation von krankheitsrelevanten Begriffen in Einzelsätzen ein anderer Ansatz verfolgt werden. Unter Verwendung der UMLS-Ontologie und dem Programm *MetaMap* wurde ein Verfahren entwickelt, das auf einer Abstraktion der gefundenen Konzepte durch ihre semantischen Felder basiert. Als krankheitsrelevante Begriffe wurden alle Konzepte gewertet, die dem semantischen Feld ‚Krankheit oder Syndrom' der Ontologie zugeordnet waren.



(a) Vorher



(b) Nachher

Abbildung 2.4 · Theoretisches Beispiel für die Zuordnung von Enzymklassen zu MESH-Begriffen. Abbildung (a) zeigt einen Auszug der MESH-Hierarchie; in jedem Knoten sind die in den dazugehörigen Dokumenten gefundenen Nennungen von Enzymklassen und die Anzahl der Nennungen vermerkt. Abbildung (b) zeigt die Veränderung nach der Auswertung: Diabetes Typ I wird die Hexokinase zugewiesen, nachdem für das Wolfram-Syndrom keine Enzymklasse mehr als dreifach gefunden wurde und die Nennungen von Hexokinase und Cytochrom-c Oxidase in den Knoten von Diabetes Typ I verschoben wurden. Die mögliche Zuordnung der Hexokinase zum weniger spezialisierten Vorfahren Diabetes Mellitus entfällt damit. Der Hypoglykämie kann nur die Glukokinase zugewiesen werden, die mit 30 Nennungen über der durchschnittlichen Anzahl von 12 Dokumenten im Knoten liegt.

Parameter für eine satzbasierte Zuordnung von Krankheiten zu Enzymklassen

Bei der Zuordnung von Krankheiten zu Enzymklassen aufgrund der Kollokation innerhalb eines Satzes spielten hauptsächlich zwei Parameter eine Rolle: Zum einen die minimale Anzahl von Sätzen mit einer gemeinsamen Nennung, zum anderen die Bewertungszahl der von *MetaMap* durchgeführten Konzeptzuordnung. Das *MetaMap*-Programm bewertet die Konfidenz der Konzeptidentifikation über das gewichtete Mittel von vier Kriterien (Zentralität der Übereinstimmung, notwendige Variationen, Abdeckung und Zusammenhalt, siehe [10]). Als realistische Untergrenze für die Bewertung ergab sich nach manueller Sichtung von 100 Sätzen und Empfehlungen des Autors (Aronson, persönliche Mitteilung) ein Wert von 650. Die Auswirkungen schrittweise stringenterer Kriterien auf die Präzision und Vollständigkeit konnten wie in Abschnitt 2.3.4 auf Seite 32 beschrieben anhand eines Textkorpus von 1 500 manuell annotierten Sätzen überprüft werden.

Negationen

Neben den beiden Parametern der minimalen Anzahl an Kollokationen und der Bewertungszahl von Konzeptzuordnungen wurde auch der Einfluss von Negationen auf die Enzym-Krankheitszuordnungen untersucht. Sätze wie

> — *. . . konnte keine Änderung der Aktivität von Enzym A bei Patienten mit Krankheit B festgestellt werden.*

führen ohne Berücksichtigung der Negation zu einer falschen Zuordnung von Enzymklasse und beschriebener Krankheit. In medizinischen Publikationen kennzeichnen die sechs Wörter *no*, *denies*, *denied*, *not*, *none* und *without* knapp 93 % aller Verneinungen [105]. Der Einfluss von Sätzen mit Negationen wurden durch die Entfernung von Sätzen mit diesen Schlüsselwörtern untersucht.

2.3.3 Semantische Kontextanalyse

Um die Genauigkeit der satzbasierten Zuordnung von Krankheitskonzepten zu Enzymklassen zu erhöhen, erfolgte eine Klassifikation von krank-

heitsrelevanten und -irrelevanten Sätzen über einen Ansatz des maschinengestützten Lernens. Bei diesem Verfahren wird ein statistisches Modell durch einen manuell annotierten Datensatz trainiert, um danach unbekannte Daten zu klassifizieren. Ein Beispiel für ein solches Verfahren ist eine *Support Vector Machine* (SVM). Diese erstellt einen binären Klassifikator, der für jeden Eigenschaften-Vektor die Klasse +1 oder −1 zurückgibt. Die Trennfunktion für die Klassen ist eine Hyperfläche; die Klasse eines Eigenschaften-Vektors ergibt sich aus der Seite des durch die Hyperfläche getrennten Raums, in der er sich befindet. Während der Trainingsphase mittels eines annotierten Datensatzes versucht die SVM die optimale Hyperfläche zu finden. Optimal bedeutet in diesem Fall ein maximaler Abstand zwischen der Hyperfläche und den der Fläche nächsten Eigenschaften-Vektoren. Bei der Maximierung liegen diese nächsten Eigenschaften-Vektoren (die *Support Vectors*) auf beiden Seiten der Hyperfläche, welche genau zwischen den Vektoren liegt [86].

Mit dem Programmpaket LIBSVM stand eine Implementierung einer SVM zur Verfügung [34]. Als Trainings- und Testdatensatz wurden 1 000 Sätze manuell annotiert, in denen in 400 Fällen ein Zusammenhang zwischen einer Enzymklasse und einer Krankheit gefunden werden konnte. Der Attributvektor jedes Satzes bestand aus der relativen Häufigkeit seiner semantischen Felder (siehe die Einleitung dieses Abschnitts auf Seite 28). Eine Verbesserung der SVM-Parameter fand durch die in der *orange*-Bibliothek implementierte 10-fach-Kreuzvalidierung statt [112].

2.3.4 Bewertung der Krankheits-Enzymzuordnungen

Vergleich mit anderen Datenbanken

Zur Überprüfung des Prinzips der automatischen Zuordnungen von Enzymklassen und Krankheiten aufgrund einer Kollokation innerhalb eines Textes erfolgte ein Vergleich mit bereits bekannten und manuell annotierten Zuordnungen. Die Proteindatenbank *Swiss-Prot* in der Version 40.22 und die OMIM-Datenbank, Stand vom Juli 2002, ermöglichten einen solchen Vergleich. Jeder OMIM-Eintrag beschreibt eine vererbbare Krankheit und enthält unter anderem ausführliche Informationen zur Krankheit, oftmals durch Fallbeschreibungen ergänzt. Die OMIM-Einträge konnten wie

normale *PubMed*-Dokumente prozessiert und nach Enzymnamen durchsucht werden (siehe Abschnitt 2.2.2 auf Seite 25).

Jede im Text identifizierte Enzymnennung wurde einfach gewertet, wobei Nennungen des empfohlenen Enzymnamens das doppelte Gewicht erhielten, eine Verwendung der EC-Nummer das dreifache. Ebenfalls dreifache Gewichtung erhielten Nennungen eines Enzymnamens im Titel eines OMIM-Dokumentes. Alle mindestens durchschnittlich gewichteten Enzymklassen galten für die Krankheit als relevant.

Die auf diese Weise generierten Zusammenhänge konnten mittels der *Swiss-Prot*-Einträge überprüft werden. Viele der *Swiss-Prot*-Datensätze enthalten manuell annotierte Verweise auf die relevanten OMIM-Daten und EC-Nummern. Diese Beziehungen wurden als Referenzdaten definiert und mit den automatisch erstellten Zuordnungen verglichen.

Vergleich mit manuell annotierten Daten

Um die Präzision und Vollständigkeit[4] der Zuordnungen bestimmen zu können, wurde ein kleiner Datensatz manuell annotiert. Präzision und Vollständigkeit sind Standardgrößen zur Evaluierung linguistischer Systeme. Die Präzision gibt den Anteil der richtigen Ergebnisse an allen erhaltenen Ergebnissen an:

$$\text{Präzision} = \frac{\text{richtig Positive}}{\text{richtig Positive} + \text{falsch Positive}}$$

Dabei steht ‚richtig Positive' für die Anzahl der richtig erstellten Zuordnungen, ‚falsch Positive' für die automatisch erstellten Zuordnungen, die sich nicht im Datensatz fanden. Die Vollständigkeit beschreibt den Anteil erhaltener richtiger Ergebnisse im Verhältnis zu allen möglichen richtigen Ergebnissen:

$$\text{Vollständigkeit} = \frac{\text{richtig Positive}}{\text{richtig Positive} + \text{falsch Negative}}$$

Die ‚falsch Negativen' stehen für Zusammenhänge im annotierten Datensatz, für die keine automatische Zuordnung vorlag. Zur Ermittlung von

[4] Im Englischen *precision* und *recall*.

Präzision und Vollständigkeit für die Zuordnung mittels Kurzzusammenfassung und MESH-Begriffen wurden die Zusammenhänge von Enzymen und medizinischen MESH-Schlagworten in 250 *PubMed*-Dokumenten manuell annotiert. Zur Bewertung der konzeptbasierten Zuordnung in Einzelsätzen erfolgte eine Annotation von 1 500 zufällig ausgewählten Sätzen mit mindestens einer Enzymnennung. Anhand der manuell annotierten Sätze konnte die Auswirkung der verschiedenen Parameter – Negationen, Mindestanzahl an Kollokationen, Bewertungszahl der Konzeptidentifikation und Einfluss des semantischen Kontexts – auf die Qualität der Zuordnung analysiert werden.

2.4 Netzwerke aus Krankheiten und Enzymklassen

2.4.1 Visualisierung der Netzwerke

Die erstellten Zuordnungen zwischen Enzymklassen und Krankheiten lassen sich als Netzwerke oder Graphen visualisieren und so durch den Benutzer einfacher auswerten, als dies über eine Textdarstellung möglich wäre. Bei Graphen handelt es sich um eine Menge von Knoten, welche durch Kanten miteinander verbunden sind. Die Knoten repräsentieren in dieser Arbeit Enzymklassen oder Krankheiten, welche aufgrund von gemeinsamen Eigenschaften durch Kanten miteinander verbunden werden. In Abbildung 2.5 finden sich Beispiele für die hier betrachteten unterschiedlichen Darstellungsweisen der Graphen:

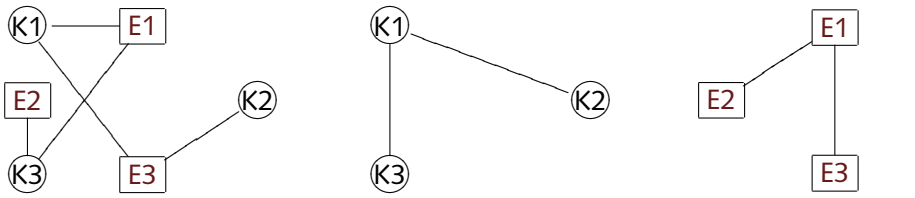
- Die Knoten im Netzwerk aus Krankheiten (Abbildung 2.5b) repräsentieren Krankheitskonzepte. Eine Kante zwischen zwei Knoten existiert hier genau dann, wenn beide Krankheiten mindestens einer gemeinsamen Enzymklasse zugeordnet sind.
- Im Enzymnetzwerk (Abbildung 2.5c) entspricht jeder Knoten einer Enzymklasse. Kanten zwischen zwei Knoten existieren genau dann, wenn beide Enzymklassen mit mindestens einer Krankheit gemeinsam assoziiert sind.
- Bei der Darstellung als sogenannter bipartiter Graph (Abbildung 2.5a) werden Krankheiten und Enzymklassen gleichzeitig als Knoten repräsentiert. In diesem Graphen existieren nur Kanten zwi-

schen Enzymen und Krankheitskonzepten, und zwar genau dann, wenn zwischen diesen beiden während der automatischen Zuordnung eine Verbindung gefunden wurde.

Für die Visualisierung der Netzwerke kamen verschiedene Programme zur Anwendung, da keine frei verfügbare Software gleichzeitig alle gestellten Anforderungen erfüllte. Zur interaktiven Analyse der Graphen bot sich die Verwendung des *Touchgraph*-Pakets an [154]. Das Programm eignet sich besonders für die Analyse der bipartiten Netzwerke. Knoten, die Enzyme repräsentieren, enthalten eine Verknüpfung mit den entsprechenden Einträgen der BRENDA-Datenbank, Krankheiten repräsentierende Knoten verweisen direkt auf die *PubMed*-Dokumente, aus denen der Begriff extrahiert wurde. Optional einblendbare Kurzzusammenfassungen (siehe unten) sowie zusätzliche Kanten zwischen Enzymknoten zur Verfolgung von Substrat-Produkt-Ketten erleichtern die Analyse. Ebenfalls optional sind zusätzliche Kanten zwischen krankheitsrepräsentierenden Knoten aufgrund einer hohen Textähnlichkeit der zugrunde liegenden *PubMed*-Dokumente.

Für eine Gesamtübersicht der MESH-Hierarchie und des Krankheitsnetzwerks eignet sich *Walrus* [75]. Die Krankheitsbegriffe der Hierarchie bilden die Knoten des Netzwerks. Kanten zwischen zwei Knoten existierten genau dann, wenn die jeweiligen Krankheiten mit mindestens einem gemeinsamen Enzym assoziiert sind. Zusätzliche Kanten des aufspannenden Baums ergeben sich durch die Vorfahren-Nachfahren-Beziehungen der MESH-Hierarchie. Für die Darstellung des Enzym- oder Krankheitsgraphen bei einer hohen Zahl von Kanten eignen sich sowohl das zur Analyse von Computernetzwerken entwickelte Programm *Otter* [72] als auch das in *Pajek* – einem Programm zur Graphenanalyse – eingebaute Visualisierungsmodul [15].

Die für die verschiedenen Visualisierungsprogramme implementierten Adapter und Ausgabemodule erlauben es, entweder ganze Netzwerke auszugeben oder sich auf Subnetzwerke, zum Beispiel über eine Eingrenzung auf bestimmte Enzyme und Krankheiten, zu beschränken.



(a) bipartites Netzwerk (b) Krankheitsnetzwerk (c) Enzymnetzwerk

Abbildung 2.5 · Beispielgraphen für Krankheiten (K) und Enzymklassen (E). Abbildung (a) zeigt ein bipartites Netzwerk, bei dem Kanten zwischen zwei Knoten für eine Zuordnung einer Krankheit zu einer Enzymklasse stehen. Abbildung (b) zeigt das gleiche Netzwerk ohne die Enzymklassen; eine Kante zwischen zwei Knoten besteht genau dann, wenn beiden Krankheiten mindestens eine Enzymklasse gemeinsam zugeordnet worden ist. Abbildung (c) zeigt den entsprechenden Graphen für Enzymklassen.

2.4.2 Visualisierung zusätzlicher Eigenschaften

Berechnung von Textähnlichkeiten

Zur Gruppierung ähnlicher Krankheiten während der Visualisierung mittels *Touchgraph* konnte die Textähnlichkeit der sie beschreibenden Dokumente oder Sätze verwendet werden. Jeweils alle ein bestimmtes Krankheitskonzept erwähnende Kurzzusammenfassungen (bei MESH-Begriffen) oder Sätze (bei Krankheitskonzepten) wurden zu einem Text zusammengefasst. Die Ähnlichkeit zwischen einem Textpaar q und d – und damit zwei Krankheiten – kann über das innere Produkt der beiden Textvektoren Q und D berechnet werden ([131], siehe auch Abschnitt 2.1.2 auf Seite 22). Mit den aus den gewichteten Termen q und d gebildeten Textvektoren

$$Q = (q_1, q_2, \ldots, q_m)$$

und

$$D = (d_1, d_2, \ldots, d_m)$$

definiert sich die Ähnlichkeit $sim(q, d)$ dann als inneres Produkt der Vektoren:

$$sim(q, d) = Q \cdot D = \sum_{j=1}^{m} q_j d_j,$$

mit m für die Anzahl der im Textkorpus verwendeten Konzepte. Da unterschiedlich große Texte eine unterschiedliche Anzahl verschiedener Terme und damit unterschiedlich lange Textvektoren haben, erfolgt abschließend eine Kosinusnormierung:

$$sim_n(q, d) = \frac{\sum_{j=1}^{m} q_j d_j}{\sqrt{\sum_{j=1}^{m} (q_j)^2 \sum_{j=1}^{m} (d_j)^2}}$$

Nach einem Vergleich aller Texte miteinander erhielten die 10 % der Krankheitsknoten mit der höchsten Textähnlichkeit eine zusätzliche Kante, welche auf die Ähnlichkeit der Krankheitsbeschreibungen hinweist.

Automatische Zusammenfassung

Zur schnelleren Einarbeitung in neue Themengebiete konnte die in Abschnitt 2.1.2 auf Seite 22 beschriebene Vektor-Repräsentation zur automatischen Erstellung einfacher Zusammenfassungen genutzt werden. Die drei Sätze mit dem jeweils höchsten Mittelwert ihrer Wort-Gewichte und mindestens einem Krankheitskonzept aus der UMLS-Ontologie bildeten zusammen mit der Konzeptdefinition der Ontologie die Zusammenfassung.

Gruppierung von Enzymen und Krankheiten

Eine hierarchische Gruppierung der Krankheiten aufgrund der zugeordneten Enzyme ermöglichte eine von der Netzwerkdarstellung abweichende Betrachtungsweise und die Suche nach interessanten Kombinationen von Krankheiten. Dazu wurde die *XCluster*-Software[5] mit den Standardeinstellungen verwendet, die ein paarweises durchschnittliches Verknüpfungsverfahren verwendet. Die resultierenden Cluster konnten mit *TreeView* [55] visualisiert werden.

[5] G. Sherlock, nicht veröffentlicht, http://genome-www.stanford.edu/~sherlock/cluster.html

2.4.3 Auswertung der Krankheits-Enzymzuordnungen

Zur Auswertung und Generierung von Hypothesen aus den erstellten Zusammenhängen zwischen Enzymen und Krankheiten kamen verschiedene Betrachtungsweisen der Netzwerke und ihrer Komponenten zur Anwendung.

Metabolische Pfade

Eine Erstellung von Hypothesen war über die Substrat-Produkt-Ketten der Enzymklassen möglich. Untersucht wurden dabei aufeinanderfolgende Enzymklassen, bei denen einem Enzym die Krankheitsannotation fehlte, die in Vorläufer- und Nachfolgeenzym vorlag (siehe Abbildung 2.6. Die fehlende Annotation kann biologisch begründet sein oder aufgrund eines nicht aus der Literatur extrahierten Zusammenhangs fehlen; sie kann aber auch Hinweise auf interessante Verbindungen liefern.

Um den Suchbereich überschaubar zu halten, beschränkten sich die untersuchten Substrat-Produkt-Ketten jeweils auf einen metabolischen Pfad. Die Daten hierfür lieferten die in der *Kyoto Encyclopedia of Genes and Genomes* (KEGG)-Datenbank enthaltenen manuell erstellten Karten, welche Verweise auf die beteiligten Enzymklassen enthalten [83]. Informationen zu Substraten und Produkten der Enzymklassen wurden ebenfalls der KEGG-Datenbank entnommen, da die metabolischen Karten auf diesen Substrat-Produktketten beruhen. Bei der Auswertung der Ergebnisse half die Abbildung der Krankheitszuordnungen auf die metabolischen Karten durch das Programm *tagKEGGMap* von Ralph Schunk (nicht veröffentlicht).

Vergleich mit Sequenzfamilien

Die Sequenzähnlichkeit der den Krankheiten zugewiesenen Enzymklassen wurden untersucht. Die Analyse basierte auf einem Vergleich aller Proteinsequenzen untereinander, die in *Swiss-Prot* (Version 41.23), und *TrEMBL* (Version 24.11) als Annotation eine EC-Nummer enthielten [21]. Die Daten beinhalteten über 270 000 Cluster basierend auf 565 000 Regionen oder Domänen aus 62 200 Aminosäuresequenzen von Enzymen (Christian aus dem Spring, persönliche Mitteilung). Von Interesse für eine genauere Untersuchung waren Cluster, die mindestens zwei Sequenzen

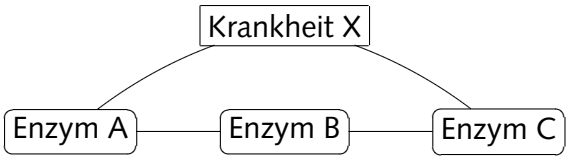


Abbildung 2.6 · Beispiel für die in metabolischen Pfaden gesuchten Dreiergruppen von Enzymen. Enzym A, B und C sind Teil eines metabolischen Pfades, dabei ist nur das mittlere Enzym B nicht mit Krankheit X assoziiert.

unterschiedlicher Enzymklassen enthielten, denen die gleiche Krankheit zugeordnet werden konnte.

Netzwerktopologie

Die Topologie der erstellten Netzwerke aus Krankheiten und Enzymen wurde anhand verschiedener Kriterien analysiert. Dies erlaubt Aussagen über die Art der Vernetzung von Krankheiten und Enzymklassen und Spekulationen über die biologischen Gründe für die Art der Vernetzungen. Zudem erlaubt die Bestimmung der Kriterien einen Vergleich mit den Eigenschaften anderer, gut untersuchter Netzwerke. Insgesamt gaben vier verschiedenen Kennzahlen Auskunft über die Struktur der beiden Netzwerke [146]:

- *Durchschnittlicher Abstand* L*:* Die Pfadlänge in einem Graphen ist die Anzahl von Kanten entlang des Pfades zwischen zwei Knoten. Dabei gibt L die durchschnittliche Länge des kürzesten Pfades zwischen allen Paaren von Knoten an, dies entspricht dem mittleren Abstand zweier beliebiger Knoten im Netzwerk. Der Abstand zweier Knoten ist definiert als die minimale Anzahl von Kanten entlang des Pfades zwischen beiden Knoten.

- *Durchmesser des Netzwerks* D*:* Der Durchmesser eines Netzwerks entspricht dem maximalen im Graphen vorkommenden Abstand zwischen zwei Knoten, also der minimalen Anzahl von Schritten mit der von einem Knoten jeder beliebige andere Knoten erreicht werden kann.

- *Gruppierungs-Koeffizient* C*:* Dieser Koeffizient beschreibt die Konnektivität des Graphen. C gibt die relative Häufigkeit an, mit der zwei Nachbarn eines beliebigen Knotens miteinander ebenfalls benachbart sind. Für einen Knoten i gibt der Gruppierungs-Koeffizient C_i das Verhältnis von existierenden Kanten T_i zwischen seinen k_i Nachbarknoten zur maximalen Anzahl möglicher Kanten $\frac{k_i(k_i-1)}{2}$ an. Der Gruppierungs-Koeffizient C ist dann als der Durchschnitt aller Knoten definiert [45]:

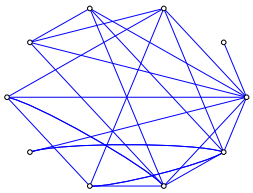
$$C = \langle C_i \rangle_i = \left\langle \frac{2T_i}{k_i(k_i-1)} \right\rangle_i$$

  In einem vollständig verknüpften Netzwerk ist $C = 1$, das heißt jeder Knoten ist mit allen anderen Knoten direkt verbunden, während es bei $C = 0$ keine Knoten mit Nachbarn gibt, die auch zueinander benachbart sind [146]. Graphen können die gleiche Anzahl an Kanten haben, aber dennoch sehr unterschiedliche Gruppierungs-Koeffizienten aufweisen.

- *Verteilung der Knotengrade* $P(k)$: Als Knotengrad bezeichnet man die Anzahl der Kanten, mit denen ein Knoten mit anderen Knoten des Netzwerks verbunden ist. Die Verteilung $P(k)$ beschreibt die Wahrscheinlichkeit, dass ein beliebiger Knoten den Grad k hat, wobei k Werte von 0 bis $n-1$, der Anzahl der restlichen Knoten im Graphen, annehmen kann. Eine Abschätzung dieser Wahrscheinlichkeitsfunktion basiert auf der relativen Häufigkeit der Knotengrade in den untersuchten Graphen. Dargestellt wird sie durch Auftragung als Punktwolke in doppelt-logarithmischen Diagrammen von $P(k)$ gegen k. Basierend auf diesen Diagrammen lassen sich Aussagen über die vermutliche Verteilungsfunktion des Graphen machen [146].

Suche nach Subgraphen

Bei der Suche nach Zusammenhängen zwischen Krankheiten aufgrund der mit ihnen assoziierten Enzyme können die Transitivitätseigenschaf-



$n = 10$
$k = 2.5$
$D = 3$
$L = 1.53$
$C = 0.28$

Abbildung 2.7 · Beispiel für die in den krankheits- und enzymbasierten Netzwerken untersuchten Eigenschaften: Ein zufällig erstellter Graph mit der Knotenzahl n, einer durchschnittlichen Konnektivität k, einem Durchmesser D und einem mittleren Abstand L bei einem Gruppierungs-Koeffizienten C.

ten[6] des Netzwerks mit Krankheitskonzepten als Knoten untersucht werden. Von Interesse waren Subgraphen mit vier Knoten, denen zur vollständigen Konnektivität genau eine Kante fehlt (siehe Abbildung 2.8). Die beiden Knoten, zwischen denen keine Kante existiert, werden dabei implizit durch die restlichen Kanten miteinander in Verbindung gebracht.

Eine Bewertung der Subgraphen erfolgte nach zwei Kriterien: Zum einen dem durchschnittlichen Knotengrad der an dem Subgraphen beteiligten Knoten, zum anderen der Anzahl von verschiedenen Subgraphen mit einer fehlenden Verbindung zwischen dem gleichen Knotenpaar. Ein geringer durchschnittlicher Knotengrad erleichtert die manuelle Auswertung der Zusammenhänge, daher wurden die Subgraphen nach ihrem Knotengrad sortiert.

2.5 Implementation

Das in dieser Arbeit entwickelte Programmpaket wurde objektorientiert in der Programmiersprache *Python* implementiert und ist betriebssystemunabhängig; ein Test fand unter *Linux* (Kernel 2.2), *Solaris* (*SunOS 5.8*) und *Windows 2000* statt. Eine Auflistung der verwendeten Module und benötigten Bibliotheken findet sich in der Anleitung des beiliegenden Programmpaketes (siehe Anhang A.7 auf Seite 138).

[6] Die Transitivität ist eine mathematische Eigenschaft binärer Beziehungen, so dass gilt: Wenn A und B sowie B und C miteinander in Beziehung stehen, dann besteht auch eine Beziehung zwischen A und C. Ein Beispiel für eine solche transitive Beziehung ist das mathematische ‚größer als'.

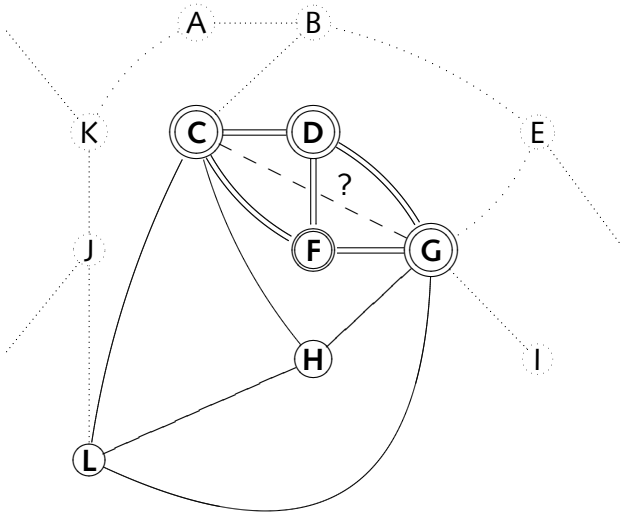


Abbildung 2.8 · Beispiel für die im Krankheitsgraphen gesuchten Subgraphen. Die mit doppelter Linie markierten Knoten C, D, F und G sind bis auf eine fehlende Kante zwischen den Krankheiten C und G vollständig konnektiert (gestrichelte Linie). Ein weiterer Subgraph mit dieser Eigenschaft findet sich mit den Knoten C, G, H und L, bei dem ebenfalls nur C und G nicht benachbart sind. Die gepunkteten Linien zeigen Verbindungen zu Knoten, die nicht Teil des gesuchten Subgraphen sind.

Für eine Zuordnung von Phrasen zu Konzepten wird eine Lizenz der UMLS-Ontologie benötigt. Dies erlaubt auch die Benutzung des *MetaMap*-Paketes erlaubt. Alternativ kann hierfür der von der *Semantic Knowledge Representation Group* zur Verfügung gestellte Server genutzt werden.

Aufbau der Datenbank

Eine lokale Speicherung sämtlicher Daten in dem relationalen Datenbanksystem MYSQL (Version 3.23.45) erleichterte die Zugriffsmöglichkeiten und die Erstellung von Zusammenhängen zwischen Enzymklassen und Krankheiten. Ein Datenbankschema findet sich im Anhang unter A.6 auf Seite 135.

3 ERGEBNISSE

> I never could make out what those damned dots meant.
>
> *Lord Randolph Churchill (1849-95)*

3.1 Enzyme und ihre Namen

Aus den Einträgen der in dieser Arbeit verwendeten Version der BRENDA-Datenbank konnte ein Lexikon von Namen für insgesamt 3 901 Enzymklassen erstellt werden (siehe Abschnitt 2.2.1 auf Seite 24). Das Lexikon enthielt die gleiche Anzahl an empfohlenen Enzymnamen und EC-Nummern sowie 17 350 Synonyme. Zusammengefasst enthielt das Lexikon somit 25 152 verschiedene Bezeichnungen für Enzyme, was inklusive der EC-Nummer einem Schnitt von mehr als sechs verschiedenen Bezeichnungen je Enzymklasse entspricht. Eine Übersicht über die zehn EC-Nummern mit der höchsten Anzahl verschiedener Namen findet sich in Tabelle 3.1. Das Enzymlexikon wurde für zwei Aufgaben benötigt: Den Aufbau eines Textkorpus (siehe die Abschnitte 2.1.1 auf Seite 15 und 3.2.1 auf Seite 47) sowie die Annotation der gesammelten Texte mit Enzymklassen, die aus Gründen der Übersichtlichkeit zuerst behandelt wird.

Tabelle 3.1 · Übersicht von EC-Nummern mit der höchsten Anzahl von Synonymen. Die Namensliste besteht aus dem empfohlenen Enzymnamen, der EC-Nummer sowie allen Synonymen mit mehr als drei Buchstaben.

Anzahl Namen	EC-Nummer	Empfohlener Enzymname
180	3.1.21.4	Type II site-specific deoxyribonuclease
67	2.7.1.37	Protein kinase
65	2.4.1.17	Glucuronosyltransferase
45	3.4.21.62	Subtilisin
40	3.2.1.96	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase
40	3.1.1.3	Triacylglycerol lipase
39	3.1.3.48	Protein-tyrosine-phosphatase
38	2.6.1.19	4-Aminobutyrate transaminase
37	2.7.1.112	Protein-tyrosine kinase
37	3.1.21.3	Type I site-specific deoxyribonuclease

3.1.1 Identifikation von Enzymnamen

In 105 630 *PubMed*-Dokumenten des Textkorpus konnte mindestens eine Enzymnennung identifiziert werden. Eine Übersicht der Enzymnamen mit den häufigsten Nennungen – unabhängig von ihrem Bezug zu Krankheiten – findet sich in Tabelle 3.2. In der überwiegenden Mehrzahl der Fälle fanden sich nur die Namen von ein bis drei unterschiedlichen Enzymklassen in einem einzelnen *PubMed*-Dokument, allerdings lagen wie in Tabelle 3.3 auf der nächsten Seite gezeigt in einigen Texten auch Auflistungen von deutlich mehr Enzymklassen vor: eine Studie zur Vorbehandlung von Tumorzellen erwähnt 25 verschiedene Enzyme in Kurzzusammenfassung und MESH-Annotation (*PubMed Identifier* (PMID) 176270, [24]).

Qualität der Namensidentifikation

Bei einer stichprobenartigen Durchsicht von 100 Kurzzusammenfassungen zur Analyse der Qualität der Identifikation von Enzymnamen konnten manuell 120 Namen annotiert werden. Davon waren mehr als 97 % durch die automatische Namensidentifikation richtig zugeordnet worden. In zwei Fällen fehlte die Annotation aufgrund der Verwendung eines Synonyms mit weniger als vier Buchstaben, welches nicht Bestandteil des ver-

Tabelle 3.2 · Übersicht der am häufigsten in einem Textkorpus aus 105 897 *PubMed*-Dokumenten gefundenen EC-Nummern und deren Anzahl an Enzymnamen in der BRENDA-Datenbank.

Anzahl Dokumente	EC-Nummer	Empfohlener Enzymname	Anzahl von Namen in BRENDA
3 495	2.7.1.37	Protein kinase	67
2 926	2.7.7.49	RNA-directed DNA polymerase	9
2 668	3.1.3.1	Alkaline phosphatase	10
2 580	3.6.1.3	Adenosinetriphosphatase	14
2 564	3.4.23.15	Renin	6
2 468	3.4.15.1	Peptidyl-dipeptidase A	20
2 339	3.4.21.4	Trypsin	14
2 324	3.1.1.7	Acetylcholinesterase	12
2 285	1.1.1.27	L-Lactate dehydrogenase	13
2 132	3.4.21.5	Thrombin	13
2 099	1.15.1.1	Superoxide dismutase	23

Tabelle 3.3 · Anzahl der unterschiedlichen Enzymnennungen in den 105 630 *PubMed*-Dokumenten des Textkorpus mit mindestens einer identifizierten Enzymnennung. In mehr als der Hälfte der Dokumente wird nur ein Enzym in der Kurzzusammenfassung oder den MESH-Begriffen erwähnt.

Anzahl Enzymnennungen	Anzahl Dokumente
1	58 066
2	24 142
3	12 487
4	5 657
5	2 851
6	1 282
7	553
8	295
9	115
10–15	171
16–25	11
	105 630

wendeten Namenslexikons war. In einem weiteren Fall war das im Text verwendete Synonym nicht Bestandteil der BRENDA-Datenbank.

Fehler bei der Identifikation von Enzymnamen

Für eine zusätzliche Fehleranalyse wurden fünfzig *PubMed*-Dokumente ausgewertet, bei denen die MESH-Schlagworte einen Enzymnamen enthielten, der nicht im Text der Kurzzusammenfassung identifiziert werden konnte (siehe Tabelle 3.4 auf der nächsten Seite):

- In 38 Fällen fand sich der Enzymname nicht in der Kurzzusammenfassung, jedoch ließ sich die MESH-Indexierung der Kurzzusammenfassung mit dem nicht gefundenen Enzymnamen in einigen Fällen nachvollziehen. In neun Publikationen bestand ein erkennbarer Zusammenhang zu in der Kurzzusammenfassung beschriebenen Methoden – zum Beispiel *rna directed dna polymerase* bei der Verwendung der RT-PCR-Methode oder *beta galactosidase* bei einer Promotorenanalyse – in acht weiteren zu einer in der Publikation besprochenen Krankheit. Bei Texten über Alkoholismus fanden sich beispielsweise Verweise auf die Alkohol-Dehydrogenase, auch wenn diese selber nicht Thema der Veröffentlichung war.

- Vier Synonyme fehlten in der verwendeten Version der BRENDA-Datenbank.

- In zwei Fällen (‚Unzureichende Ausprägung') waren die im Text gefundenen Namen nicht ausreichend für eine eindeutige Zuordnung.

- In zwei weiteren Fällen stand der gefundene MESH-Begriff nicht für einen Enzymnamen. Ein Beispiel hierfür ist *topical administration* als MESH-Schlagwort, wobei *topical* zugleich als Synonym für EC-Nummer 3.4.21.5, Thrombin, im Enzymlexikon steht. In einem Fall stimmte die Enzymklasse des MESH-Begriffs und die des im Text verwendeten Synonyms nicht überein: Der MESH-Begriff *map kinase* hat im Enzymlexikon die EC-Nummer 2.7.1.123, die mögliche Abkürzung MAPK findet sich unter EC-Nummer 2.7.1.37.

Tabelle 3.4 · Fehleranalyse der automatischen Erkennung von Enzymnamen. Gezeigt ist eine Untersuchung von 50 Kurzzusammenfassungen, eine Erläuterung zu den Fehlerbegründungen findet sich im Text.

Fehlerbegründung	Anzahl
Enzymname nicht im Text erwähnt	38
——— davon 9 mit Bezug zur Methode	
——— davon 8 mit Bezug zu einer Krankheit	
Im Text verwendeter Name nicht in der BRENDA-Datenbank	4
MESH-Schlagwort kein Enzymname	2
Unzureichende Ausprägung des Enzymnamens im Text	2
Fehler in der *Token*-Bildung	1
Falsche Schreibweise des Enzymnamens im Text	1
Synonym im Text und MESH-Schlagwort mit unterschiedlicher EC-Nummer	1
Fehlende Unterscheidung Singular / Plural	1

3.1.2 Levenshtein-Distanz in Enzymnamen

Die durchgeführte Berechnung der Levenshtein-Distanz von Einträgen des Enzymnamenlexikons zueinander sollte einen Anhaltspunkt geben, inwieweit alternative Schreibweisen oder leicht von den Einträgen abweichende Enzymnennungen in Texten identifiziert werden können (siehe Abschnitt 2.2.2 auf Seite 25).

Bei einer Untersuchung der Namen von Enzymklassen mit unterschiedlicher EC-Nummer finden sich 2 050 EC-Nummern mit einer maximalen Levenshtein-Distanz von eins. Davon haben knapp 200 Enzymnamen eine unterschiedliche Hauptklasse, deutlich mehr als die bei einer Levenshtein-Distanz von null gefundenen 85 identischen Enzymnamen mit unterschiedlicher Hauptklasse. Einige der Namen, die mit nur einer Insertion, Deletion oder Substitution ineinander überführt werden können, sind in Tabelle 3.5 aufgelistet. Bei einer Levenshtein-Distanz von zwei steigt die Zahl auf 3 450 einander ähnliche Enzymnamen mit unterschiedlichen EC-Nummern. Aufgrund der hohen Anzahl ähnlicher Enzymnamen basierte die Identifikation in Texten nur auf vollständiger Übereinstimmungen mit Einträgen des Enzymlexikons – also solchen mit einer Levenshtein-Distanz von null.

Tabelle 3.5 · Beispiele für Enzymnamen mit unterschiedlicher EC-Nummer und einer Levenshtein-Distanz von eins. Unterschiede zwischen den Enzymnamen sind mit Fettdruck gekennzeichnet.

Enzym A	EC-Nummer A	Enzym B	EC-Nummer B
fad synthetase	2.7.7.2	**n**ad synthetase	6.3.1.5
ribonuclease **p**h	2.7.7.56	ribonuclease h	3.1.26.4
de**s**oxyribonuclease	3.1.21.1	deoxyribonuclease	4.2.99.18
triokinase	2.7.1.28	t**h**iokinase	6.2.1.3
lac**t**ase	3.2.1.23	laccase	1.10.3.2

3.2 Aufbau und Prozessierung des Textkorpus

3.2.1 Erstellung eines Textkorpus

Das Enyzmlexikon diente zusammen mit dem kontrollierten Vokabular der MESH-Begriffe zum Aufbau eines Textkorpus von Kurzzusammenfassungen aus der *PubMed*-Datenbank. Über Suchanfragen aus je einem krankheitsrelevanten MESH-Begriff sowie den Namen einer Enzymklasse erfolgte ein Abruf von insgesamt 105 897 *PubMed*-Dokumenten. Abbildung 3.1 zeigt die Anzahl abgerufener Dokumente im Vergleich zum Publikationsjahr. Die genaue Identifikation von Enzymnamen und deren EC-Nummer innerhalb der abgerufenen Texte erfolgte wie in Abschnitt 2.2.2 beschrieben und resultierte in der automatischen Annotation von knapp 200 000 Nennungen bekannter Enzymnamen und Synonyme im vorliegenden Textkorpus.

3.2.2 Lexikalische Analyse der Texte

Alle vorliegenden Texte wurden wie in Abschnitt 2.1.2 beschrieben lexikalisch analysiert und dabei in *Token*, Wörter und Wortstämme zerlegt sowie mit Hilfe des Programms *MetaMap* Konzepten zugeordnet. Durch die Umwandlung von *Token* in Wörter und der Filterung mittels einer Stoppwortliste (siehe Abschnitt 2.1.2) reduzierte sich die Gesamtgröße des Textkorpus um etwa 40 %. Gleichzeitig verkleinerten sich dadurch die zur Beschreibung der Dokumente notwendigen Vektoren, was den Vergleich der Texte untereinander und die automatische Erstellung von Kurzzusammenfassungen erleichterte (siehe Abschnitte 2.4.2 und 2.4.2 auf Seite 36).

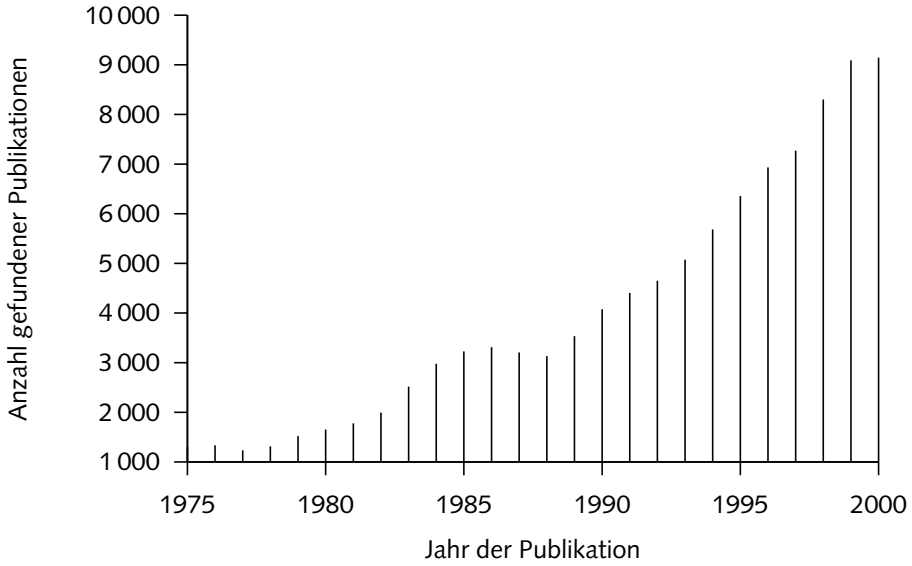


Abbildung 3.1 · Anzahl der mit Suchanfragen nach Enzymnamen und krankheitsrelevanten MESH-Begriffen gefundenen *PubMed*-Dokumente in Abhängigkeit des Publikationsjahres.

Wie in Abbildung 3.2 zu sehen kommt es aufgrund der Filterung zu einer Veränderung der relativen Worthäufigkeiten, die zur Gewichtung von Wörtern verwendet wird. Gezeigt sind die Term- und Dokumentfrequenzen für *Token* und Wörter anhand eines zufällig zusammengestellten Beispielkorpus von 100 Kurzzusammenfassungen. Die üblichen Stoppwörter wie ‚*the*', ‚*in*' oder ‚*was*' fanden sich mit hoher Termfrequenz in fast allen Dokumenten. Zusätzlich enthielten mehr als 25 % der Kurzzusammenfassungen auch *Token* wie ‚*patients*', ‚*activity*' oder ‚*results*', dafür aber mit niedrigerer Frequenz innerhalb der einzelnen Dokumente. Diese Worte spiegeln die Wortverteilung des aufgebauten Textkorpus wieder und waren Bestandteil der erweiterten *PubMed*-Stoppwortliste. In Folge der Filterung erhalten für einen Vergleich der Texte interessante Begriffe wie ‚Insulin' eine höhere relative Termfrequenz und dadurch ein höheres Gewicht.

Eine zusätzliche Reduktion des Vokabulars um 10 % erfolgte durch die Rückführung von Worten auf ihre Stammformen unter Verwendung des *Porter Stemmer*-Algorithmus (siehe Tabelle 3.6). Dieser ersetzt verschiede-

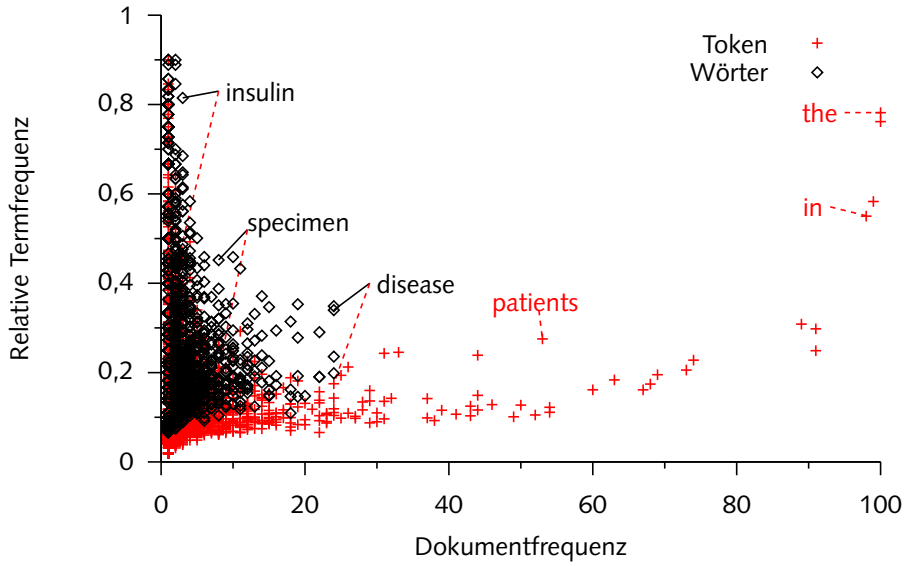


Abbildung 3.2 · Term- und Dokumentfrequenzen bei der Verwendung von *Token* und Wörtern anhand eines Testkorpus von 100 zufällig ausgewählten Kurzzusammenfassungen. Durch die Filterung von *Token* wie ‚*the*' oder ‚*in*' anhand einer Stoppwortliste ändern sich die restlichen Termfrequenzen.

ne Flektionen eines Wortes – unter anderem durch Entfernung gebräuchlicher Suffixe – durch die entsprechende Stammform. Neben der Verringerung des Vokabulars reduzierte sich auch die durchschnittliche Anzahl von Termen je Kurzzusammenfassung von 76 Worten auf 71 Stammformen. Die beste Datenreduktion zeigte sich bei der Verwendung von Konzepten. Durch die Zusammenfassung verschiedener Bezeichnungen zu einer Repräsentation verringerte sich das zur Beschreibung des Textkorpus notwendige Vokabular auf 20 % der verwendeten Worte. Die Anzahl von Termen je Kurzzusammenfassung konnte auf 42 Konzepte reduziert werden.

Terme mit hoher Dokumentfrequenz

Eine Übersicht der zehn häufigsten *Token*, Wörter, Stammformen und Konzepte des Textkorpus findet sich in Tabelle 3.7. Sämtliche aufgeführte *Token* sind Bestandteil der von der *PubMed*-Datenbank bereitgestellten Stoppwortliste. Die restlichen Spalten weisen viele Begriffe aus dem

Tabelle 3.6 · Datenreduktion durch die Filterung von *Token*, die Rückführung von Wörtern auf ihre Stammformen sowie die Zuweisung von Konzepten zu Wörtern oder Phrasen in Kurzzusammenfassungen.

	Gesamtzahl im Textkorpus	Größe des Wörterbuchs	Durchschnittliche Anzahl je Kurzzusammenfassung
Alle *Token*	20 632 662	301 711	207,8
Nur Wörter	12 620 000	270 000	76,2
Nur Stammformen	12 620 000	240 000	71,2
Nur Konzepte	7 600 000	49 000	42,0

klinischen und therapeutischen Umfeld auf, was auf einen hohen Anteil medizinischer Publikationen im Textkorpus hinweist.

3.3 Zuordnung von Krankheiten zu Enzymklassen

Die Zuordnung von Krankheitsbegriffen zu Enzymklassen basierte entweder auf der Identifikation von Enzymnamen in Kurzzusammenfassungen und der Annotation mit MESH-Begriffen, oder aber auf einer Kollokation von Enzymnamen und krankheitsrelevanten Konzepten der UMLS-Ontologie innerhalb eines Satzes (siehe die Abschnitte 2.3.1 auf Seite 26 und 2.3.2 auf Seite 28).

3.3.1 Überprüfung des Verfahrens mittels der OMIM- und *Swiss-Prot*-Datenbanken

Die in Abschnitt 2.3.4 auf Seite 31 beschriebene Auswertung der OMIM- und *Swiss-Prot*-Datenbanken sollte einen Anhaltspunkt liefern, inwieweit sich aufgrund der Kollokation von Enzymklassen und Krankheiten innerhalb eines Textes richtige Zuordnungen erstellen lassen.

In der *Swiss-Prot*-Datenbank fanden sich 1 460 manuell erstellte Verknüpfungen zwischen 636 Proteinen mit einer EC-Nummer und 1 422 Einträgen in der OMIM-Datenbank. Die eigene, automatische Auswertung der OMIM-Datenbank resultierte in 3 328 Zuordnungen zwischen 630 Enzymklassen und 2 366 OMIM-Einträgen. Von den gefundenen Zuordnungen stimmten 1 004 mit den manuellen Annotationen der *Swiss-Prot*-

Tabelle 3.7 · Übersicht der zehn häufigsten *Token*, Wörter, Stammformen und Konzepte des Textkorpus.

Rang	Token	Wörter	Stammformen	Konzepte
1	of	effect	effect	High
2	the	disease	protein	Adjudication
3	and	protein	differ	Therapeutic procedure
4	in	two	diseas	Role
5	to	significant	determin	Therapeutic aspects
6	with	normal	control	Normal
7	a	increase	suggest	Two
8	was	role	signific	Utilization
9	for	treatment	enzym	Disease
10	were	similar	two	Control

Datenbank überein. Eine aufgrund der unerwartet niedrigen Übereinstimmung durchgeführte Überprüfung von fünfzig eigenen, in der *Swiss-Prot*-Datenbank nicht annotierten Zuordnungen führte zu dem Ergebnis, dass 41 der als falsch-positiv markierten Ergebnisse richtig waren und in *Swiss-Prot* fehlten. Selbst eine sorgfältig manuell erstellte Datenbank wie *Swiss-Prot* könnte demnach von der Erweiterung durch automatische Verfahren der Informationsextraktion profitieren.

3.3.2 Zuordnung von Krankheiten zu Enzymklassen mittels Medical Subject Headings

In den 105 897 *PubMed*-Dokumenten des erstellten Textkorpus fanden sich in 94 730 Fällen mindestens ein krankheitsrelevanter MESH-Begriff sowie ein Enzymname innerhalb eines Dokuments. Von den 3 901 EC-Nummern in der BRENDA-Datenbank konnten aufgrund dieser Kollokationen 849 Enzymklassen mit einem oder mehreren Krankheitsbegriffen in Verbindung gebracht werden. Insgesamt gab es mehr als 35 000 unterscheidbare Zuordnungen zwischen Enzymklassen und den 1 900 im Textkorpus vorhandenen unterschiedlichen MESH-Krankheitsbegriffen, was einem Mittel von mehr als vierzig Zuordnungen je Enzymklasse entspricht (siebe Tabelle 3.8).

Tabelle 3.8 · Anzahl der *PubMed*-Dokumente mit der Zahl an Zuordnungen zwischen Enzymklassen und Krankheit unter Verwendung der MESH-Begriffe.

Anzahl der Zuordnungen	Anzahl *PubMed*-Dokumente	Summe der Zuordnungen
1	53 947	53 947
2	21 426	42 852
3	10 526	31 578
4	4 608	18 432
5	2 373	11 865
6	1 020	6 120
7	422	2 954
8	212	1 656
9	78	702
10+	118	1 324
	94 730	171 430

Überprüfung der Zuordnungen von Krankheiten zu Enzymklassen

Eine manuelle Annotation von 250 *PubMed*-Dokumenten auf einen Zusammenhang zwischen den identifizierten Enzymklassen und den vorgefundenen MESH-Begriffen gab genauere Auskunft über die Zuverlässigkeit der Methode. Als richtig positiv galt hierbei eine Zuordnung, in der:

- die im Dokument erwähnte Enzymklasse richtig identifiziert werden konnte,
- bei der die zugeordnete Annotation mit einem krankheitsrelevanten MESH-Begriff maximal zwei Ebenen in der MESH-Hierarchie von der manuellen Annotation des *PubMed*-Dokumentes entfernt war,
- und bei der innerhalb der Kurzzusammenfassung ein Zusammenhang zwischen identifizierter Enzymklasse und Krankheitsbegriff erkennbar war.

Als richtig negativ galten *PubMed*-Dokumente, in denen trotz gemeinsamer Nennung eines Enzymnamens und eines krankheitsrelevanten MESH-Begriffs keine Zuordnung vorgenommen wurde und ein solcher Zusammenhang im Dokument auch nicht erkennbar war. Dies war meist bei

der Filterung aufgrund einer nicht ausreichenden Anzahl von Nennungen dieser Zuordnung im Textkorpus der Fall.

Bei der ausschließlichen Verwendung von Dokumenten mit genau einem krankheitsrelevanten MESH-Begriff und einer identifizierten Enzymklasse konnte eine Präzision von 76 % bei einer Vollständigkeit von 80 % erreicht werden. Bei der Verwendung aller Dokumente und damit auch der Zuordnung von mehreren Krankheiten und Enzymklassen innerhalb eines *PubMed*-Dokumentes reduzierte sich die Präzision auf 52 % bei einer Vollständigkeit von 88 %. Aufgrund dieser niedrigen Präzision eigneten sich die Zuordnungen über Kollokationen in Kurzzusammenfassungen nicht für eine direkte Übernahme in die BRENDA-Datenbank.

3.3.3 Zuordnungen von Krankheiten zu Enzymklassen durch Konzepte

Die Auswertung von Kollokationen in kleinen Textabschnitten wie Sätzen erhöht die Präzision der extrahierten Zusammenhänge gegenüber der Auswertung von Kurzzusammenfassungen (siehe Abschnitt 2.3.2 auf Seite 28 und [50]). Die Zuordnung von Krankheitskonzepten zu Enzymklassen erfolgte aufgrund von Kollokationen in 48 164 Sätzen des Textkorpus, in denen mindestens ein krankheitsrelevantes Konzept und ein Enzymname identifiziert werden konnte. Die Qualität dieser satzbasierten Zuordnung wurde mittels eines manuell annotierten Textkorpus von 1 500 zufällig ausgewählten Sätzen verifiziert. Der Testkorpus enthielt 273 Sätze mit mindestens einem Zusammenhang zwischen einem Krankheitskonzept und einer Enzymklasse, in 80 Sätzen fand sich eine Kollokation, ohne dass ein Zusammenhang zwischen beiden erkennbar war. Insgesamt konnten 430 Zuordnungen – davon 384 unterschiedliche – annotiert werden. Dies entspricht einem Verhältnis von einem aus vier Sätzen mit Krankheitsinformationen und einer Enzymnennung.

Für die Bewertung der automatischen Zuordnung anhand von Präzision und Vollständigkeit sind zwei verschiedene Bewertungskriterien relevant:

- die richtige Identifikation jeder einzelnen Kollokation von Enzymklassen und Krankheitskonzepten im Textkorpus

- die richtige Zuordnung einer Enzymklasse zu einem Krankheitskonzept. Ein richtige Zuordnung entspricht in diesem Fall der Annotation von Enzymklasse A mit Krankheitskonzept B, unabhängig von der Zahl der richtig gefundenen Kollokationen von A und B im Textkorpus.

Tabelle 3.9(a) zeigt den Unterschied zwischen den beiden Bewertungsverfahren. Bei einer Bewertungszahl der Konzeptzuordnung durch *MetaMap* von 650 (siehe Abschnitt 2.1.2 auf Seite 19) konnten 81,4 % aller Kollokationen gefunden werden – und damit 84,8 % der im annotierten Textkorpus vorliegenden Zusammenhänge. Die Präzision lag bei beiden Bewertungen bei etwa 82 %.

Für die Zielsetzung dieser Arbeit war die Erkennung eines Zusammenhangs zwischen Enzymklasse und Krankheit wichtiger als die Identifikation jeder einzelnen Kollokation. Alle weiteren Nennungen von Präzision und Vollständigkeit beziehen sich daher auf die Erkennung der 384 unterschiedlichen Zuordnungen der 1 500 annotierten Sätze. Bei der Auswertung der satzbasierten Zuordnungen werden die Begriffe Krankheit und Krankheitskonzept synonym verwendet.

Stringentere Kriterien

Die Einführung einer Mindestbewertungszahl für Konzepte führte mit steigendem Mindestwert bei beiden Arten der Bewertung zu einer Verbesserung der Präzision um 3 %. Dies ging allerdings zu Lasten einer um 15 % niedrigeren Vollständigkeit (Tabelle 3.9(a)). Bei höheren Mindestbewertungszahlen wurden mehr richtige als falsche Zuordnungen entfernt, was zu einer sinkenden Präzision führte.

Durch eine Entfernung von Zuordnungen unter einer Mindestzahl von gefundenen Kollokationen ließ sich die Präzision um weitere 4 % steigern, dies war jedoch mit einem zusätzlichen Verlust von 14 % in der Vollständigkeit verbunden (Tabelle 3.9(b)).

Die Entfernung von Sätzen mit einem Negationswort führte durchgängig zu einer Verbesserung der Präzision um 2–3 % bei einer um 3 % verringerten Vollständigkeit. Bei einer großen Mindestzahl von Kollokationen verschlechterte sich die Vollständigkeit bei der Entfernung von Sätzen mit einer Negation nur noch minimal. Abbildung 3.3 auf Seite 56 verdeutlicht die in Tabelle 3.9(b) gezeigten Ergebnisse.

Tabelle 3.9 · Präzision und Vollständigkeit der Zuordnung von krankheitsrelevanten Konzepten zu Enzymklassen in Prozent. Die jeweils erste Zahl bezieht sich auf die richtige Erkennung jeder einzelnen Kollokation von Enzym und Krankheit, die zweite Zahl auf die richtige Zuordnung innerhalb des gesamten Textkorpus. Tabelle (a) zeigt die Abhängigkeit von Präzision und Vollständigkeit der Zuordnung von einer Filterung aufgrund der Mindestbewertungszahl der Konzepte unabhängig von der Mindestzahl gefundener Kollokationen, Tabelle (b) die Veränderung mit der Mindestzahl an gefundenen Kollokationen bei einer Mindestbewertungszahl von 800. Bei der Bewertung mit Satznegationen wurden Zuordnungen auch aufgrund von Sätzen mit einer Negation erstellt, bei der Bewertung ohne Negation erfolgte eine vorherige Entfernung dieser Sätze.

(a) Präzision und Vollständigkeit in Abhängigkeit der Mindestbewertungszahl der Konzeptzuordnung ohne Mindestzahl an Kollokationen

Mindest-	mit Negationen		ohne Negationen	
bewertungszahl	Präzision	Vollständigkeit	Präzision	Vollständigkeit
650	81.4 / 82.1	81.4 / 84.8	83.7 / 84.3	76.3 / 80.2
700	81.8 / 82.5	69.2 / 73.3	84.5 / 85.0	64.9 / 69.0
750	81.6 / 82.4	67.1 / 71.1	84.4 / 84.9	62.8 / 66.7
800	84.3 / 85.2	65.0 / 69.9	87.5 / 88.1	60.6 / 65.6
850	83.8 / 84.6	58.8 / 63.2	87.4 / 87.8	54.5 / 59.0
900	84.3 / 85.4	47.7 / 51.7	86.6 / 87.2	44.2 / 47.7
950	81.5 / 83.6	34.5 / 38.9	84.0 / 84.8	32.4 / 35.8
1 000	81.4 / 83.7	33.1 / 37.4	84.1 / 84.9	31.1 / 34.3

(b) Präzision und Vollständigkeit in Abhängigkeit von der Mindestzahl an Kollokationen bei einer Mindestbewertungszahl von 800

Mindestzahl	mit Negationen		ohne Negationen	
an Kollokationen	Präzision	Vollständigkeit	Präzision	Vollständigkeit
1	84.3 / 85.2	65.0 / 69.9	87.5 / 88.1	60.6 / 65.6
2	87.5 / 88.8	52.5 / 55.6	90.7 / 91.6	49.6 / 52.9
3	87.7 / 89.4	42.1 / 43.8	90.3 / 91.4	39.9 / 41.2
4	88.0 / 89.9	39.4 / 40.7	89.8 / 90.9	37.8 / 39.5
5	87.9 / 90.1	35.1 / 36.2	90.0 / 91.3	33.8 / 35.3
6	87.9 / 90.3	33.2 / 34.0	90.1 / 91.5	31.9 / 33.1
7	88.1 / 90.6	31.9 / 32.5	90.4 / 92.0	30.6 / 31.6
8	88.2 / 90.1	30.3 / 30.7	90.7 / 92.4	28.9 / 29.7
9	88.4 / 91.3	28.7 / 28.9	91.1 / 92.9	27.3 / 27.9
10	88.0 / 91.0	27.6 / 27.6	90.7 / 92.6	26.3 / 26.7

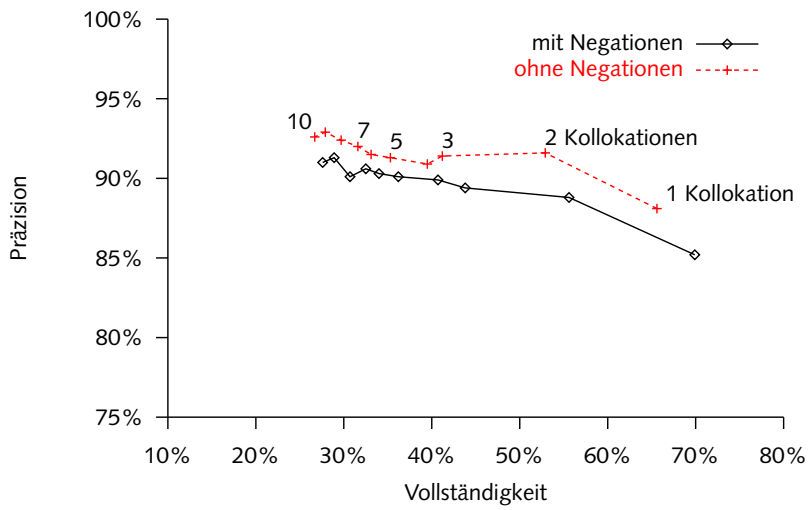


Abbildung 3.3 · Veränderung von Präzision und Vollständigkeit bei der satzbasierten Zuordnung von Enzymklassen zu Krankheitskonzepten. Gezeigt ist die Verbesserung der Präzision durch die Filterung von Sätzen mit Negationswörtern in Abhängigkeit von der Mindestzahl an gefundenen Kollokationen (siehe Tabelle 3.9(b)).

Beschränkung auf eindeutige Zuordnungen von Krankheiten zu Enzymklassen

Tabelle 3.10 auf der nächsten Seite zeigt die erreichten Werte bei der Verwendung von Sätzen mit nur einer Enzymklasse und genau einem Krankheitskonzept. Durch die Beschränkung auf solche eindeutigen Kollokationen reduzierte sich die Anzahl der auswertbaren Sätze um knapp 50 % auf 26 892. Die Unterschiede in Präzision und Vollständigkeit im Vergleich zur Verwendung aller Kollokationen zeigt Abbildung 3.4.

Nur bei einer sehr niedrigen Vollständigkeit lässt sich durch die Verwendung eindeutiger Kollokationen die Präzision leicht steigern. Aufgrund der nur geringen Verbesserung und den deutlichen Verlusten an Vollständigkeit wurden im weiteren alle Sätze verwendet, unabhängig davon wie viele Enzymklassen und Krankheitskonzepte darin vorkamen.

Tabelle 3.10 · Präzision und Vollständigkeit bei der satzbasierten Zuordnung von krankheitsrelevanten Konzepten zu Enzymklassen in Prozent. Im Unterschied zu Tabelle 3.9 wurden nur Sätze mit genau einer Enzymklasse sowie einem Krankheitskonzept verwendet. Tabelle (a) zeigt die Abhängigkeit von Präzision und Vollständigkeit der Zuordnung von einer Filterung von Konzepten aufgrund der Mindestbewertungszahl der Konzepte ohne Berücksichtigung der Mindestzahl an Kollokationen, Tabelle (b) die Veränderung mit der Mindestzahl an gefundenen Kollokationen bei einer Mindestbewertungszahl von 800. Bei der Bewertung mit Satznegationen wurden Zuordnungen auch aufgrund von Sätzen mit einer Negation erstellt, bei der Bewertung ohne Negation erfolgte eine vorherige Entfernung dieser Sätze.

(a) Präzision und Vollständigkeit in Abhängigkeit von der Mindestbewertungszahl ohne Mindestzahl an Kollokationen

Mindest-	mit Negationen		ohne Negationen	
bewertungszahl	Präzision	Vollständigkeit	Präzision	Vollständigkeit
650	78.3 / 77.6	28.2 / 28.6	81.5 / 80.5	26.0 / 26.4
700	80.1 / 81.1	24.9 / 26.1	83.6 / 84.2	23.3 / 24.3
750	80.1 / 81.2	23.8 / 24.9	84.0 / 84.6	22.5 / 23.4
800	81.1 / 81.4	23.1 / 24.0	85.3 / 85.2	21.7 / 22.8
850	80.2 / 80.0	20.6 / 21.8	84.7 / 83.7	19.3 / 20.4
900	83.3 / 83.6	17.4 / 18.5	87.3 / 86.5	16.6 / 17.6
950	82.0 / 83.1	13.4 / 14.9	85.5 / 84.9	12.6 / 13.7
1 000	82.5 / 83.6	12.6 / 14.0	86.3 / 85.7	11.8 / 12.8

(b) Präzision und Vollständigkeit in Abhängigkeit von der Mindestzahl an Kollokationen bei einer Mindestbewertungszahl von 800

Mindestzahl	mit Negationen		ohne Negationen	
an Kollokationen	Präzision	Vollständigkeit	Präzision	Vollständigkeit
1	81.1 / 81.4	23.1 / 24.0	85.3 / 85.2	21.7 / 22.8
2	85.2 / 86.1	20.1 / 20.1	88.7 / 89.0	19.0 / 19.7
3	85.5 / 86.6	17.4 / 17.6	88.6 / 88.8	16.6 / 17.0
4	87.5 / 88.8	16.9 / 17.0	89.5 / 90.0	16.1 / 16.4
5	89.2 / 91.2	15.5 / 15.8	91.8 / 92.7	15.0 / 15.5
6	88.7 / 90.7	14.7 / 14.9	91.3 / 92.3	14.2 / 14.6
7	91.5 / 94.1	14.4 / 14.6	92.8 / 94.0	13.9 / 14.3
8	92.7 / 95.7	13.6 / 13.6	94.2 / 95.5	13.1 / 13.3
9	92.7 / 95.7	13.6 / 13.6	94.2 / 95.6	13.1 / 13.3
10	92.6 / 95.6	13.4 / 13.4	94.1 / 95.5	12.8 / 13.0

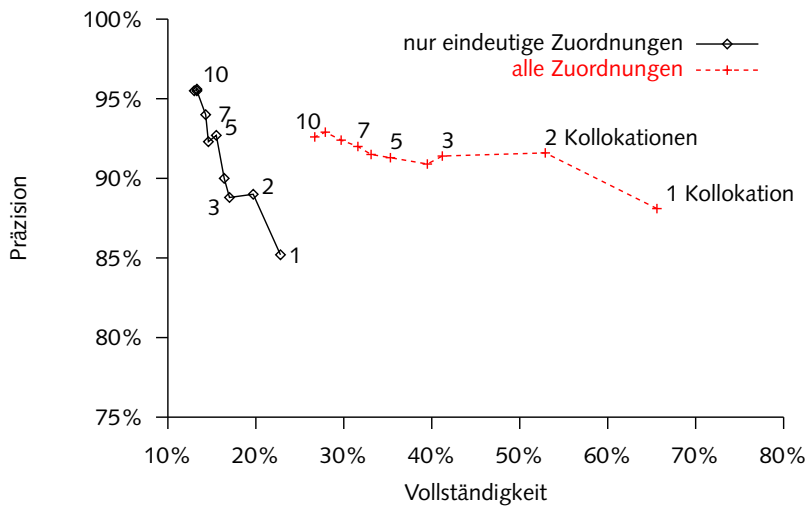
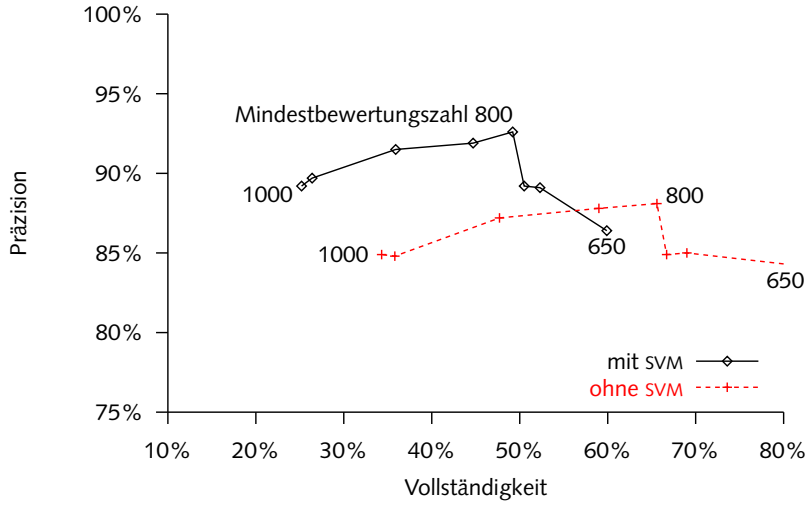


Abbildung 3.4 · Veränderung von Präzision und Vollständigkeit bei der satzbasierten Zuordnung von Enzymklassen zu Krankheitskonzepten mit der Mindestzahl gefundener Kollokationen. Die Beschränkung auf eindeutige Zuordnungen verringert die Vollständigkeit deutlich (siehe Tabellen 3.9(b) und 3.10(b)).

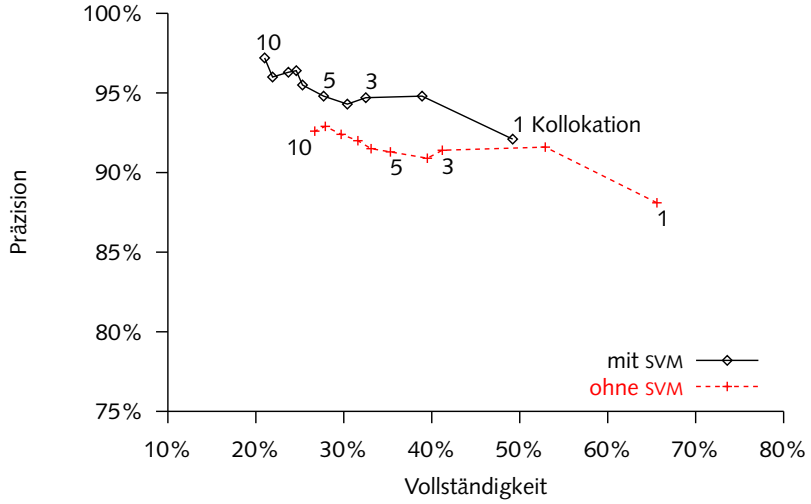
Berücksichtigung des semantischen Kontexts

Eine weitere Steigerung an Präzision bei der Zuordnungen von Enzymklassen zu Krankheitskonzepten konnte durch die Berücksichtigung des semantischen Kontexts der Konzepte erreicht werden. Zur Klassifikation von Sätzen aufgrund ihrer semantischen Felder wurde eine *Support Vector Machine* anhand eines manuell annotierten Datensatzes trainiert (siehe Abschnitt 2.3.3 auf Seite 30). Die semantischen Felder mit der höchsten Relevanz zur Unterscheidung beider Arten von Sätzen aufgrund der Bewertung von Attributen durch die *orange*-Bibliothek sind in Tabelle 3.11 aufgelistet.

Die Auswirkung der Verwendung der SVM-Klassifikation als Filter zur Analyse des semantischen Kontexts zeigt Tabelle 3.12 auf Seite 61. Die Verwendung des Filters ermöglichte unabhängig von anderen Parametern eine Verbesserung der Zuordnungspräzision bis zu einer Vollständigkeit von 50 % (siehe Abbildung 3.5 auf der nächsten Seite). Bei ansonsten gleichen Einstellungen der Filter sank die Vollständigkeit um 15-20 % bei einer Steigerung der Präzision um circa 4 %.



(a) Präzision und Vollständigkeit in Abhängigkeit der Mindestbewertungszahl der Konzeptzuordnung ohne Mindestzahl an Kollokationen



(b) Präzision und Vollständigkeit in Abhängigkeit von der Mindestzahl an Kollokationen bei einer Mindestbewertungszahl von 800

Abbildung 3.5 · Änderung der Präzision und Vollständigkeit der satzbasierten Zuordnung bei Verwendung des SVM-Filters. Bei gleichbleibender Vollständigkeit führt die Verwendung des Filters zu einer Verbesserung der Präzision sowohl bei Änderung der minimalen Bewertungszahl (Abbildung (a)) als auch bei der Mindestzahl an Kollokationen (Abbildung (b)). In beiden Fällen waren keine Sätze mit einer Negation enthalten.

Tabelle 3.11 · Semantische Felder mit der höchsten Relevanz für die Unterscheidung zwischen krankheitsrelevanten und irrelevanten Sätzen des Trainingsdatensatzes.

Rang	Feld
1	T169 (Functional Concept)
2	T101 (Patient or Disabled Group)
3	T170 (Intellectual Product)
4	T025 (Cell)
5	T121 (Pharmacologic Substance)
6	T116 (Amino Acid, Peptide, or Protein)
7	T023 (Body Part, Organ, or Organ Component)
8	T100 (Age Group)
9	T045 (Genetic Function)
10	T060 (Diagnostic Procedure)

3.4 Auswertung der Krankheits-Enzymzuordnungen

Für die weitere Auswertung wurden die Datensätze mit stringenten Parametern erzeugt: eine Mindestbewertungszahl durch *MetaMap* von 800, mindestens zwei Kollokationen nach der Entfernung von Sätzen mit einer Negation und der Verwendung des SVM-Filters. Mit diesen Parametern wurde für den Textkorpus aus 1 500 Sätzen eine Präzision von 94 % erreicht. Damit konnten die Aufgabenstellung voll erfüllt werden, der BRENDA-Enzymdatenbank krankheitsrelevante Informationen mit hoher Präzision hinzuzufügen. Insgesamt konnten so 524 Enzymklassen 1 409 Krankheitskonzepten zugeordnet werden. Die mittlere Zahl der Krankheitskonzepte je Enzymklasse reduzierte sich von über 40 bei der Verwendung der MESH-Begriffe auf 10,3 bei der konzeptbasierten Zuordnung[1].

Wahlweise können nach Bedarf des Benutzers des entwickelten Programms auch weniger stringente Parameter gewählt werden, um beispielsweise bei niedrigerer Präzision mehr Zuordnungen zu erhalten. So kann bei einer Mindestzahl von Kollokationen von eins – also der Verwendung aller gemeinsamer Nennungen – die Vollständigkeit von 39 % auf 50 % gesteigert werden. Dabei verringert sich die Präzision nur geringfügig um 2 % auf 92 %. Alle weiteren Ergebnisse beziehen sich jedoch auf

[1] Veränderungen der Parameter können zu abweichenden Ergebnissen bei der Betrachtung und Auswertung der Zuordnungen führen.

Tabelle 3.12 · Präzision und Vollständigkeit der Zuordnung von krankheitsrelevanten Konzepten zu Enzymklassen mit und ohne Filterung durch eine *Support Vector Machine* in Prozent. Die jeweils erste Zahl bezieht sich auf die richtige Erkennung jeder einzelnen Kollokation von Enzym und Krankheit, die zweite Zahl auf die Erkennung der richtigen Zuordnung innerhalb des gesamten Textkorpus. Tabelle (a) zeigt die Abhängigkeit von Präzision und Vollständigkeit der Zuordnung von einer Filterung von Konzepten aufgrund der Mindestbewertungszahl der Konzepte ohne Berücksichtigung der Mindestzahl an Kollokationen, Tabelle (b) die Veränderung mit der Mindestzahl an gefundenen Kollokationen bei einer Mindestbewertungszahl von 800. Bei der Bewertung der Zuordnungen unter Verwendung des SVM-Filters wurden alle Sätze entfernt, die das Modell als nicht krankheitsrelevant klassifizierte. In beiden Fällen waren keine Sätze mit Negationen enthalten.

(a) Präzision und Vollständigkeit in Abhängigkeit der Mindestbewertungszahl der Konzeptzuordnung ohne Mindestzahl an Kollokationen

Mindest-bewertungszahl	ohne SVM		mit SVM	
	Präzision	Vollständigkeit	Präzision	Vollständigkeit
650	83.7 / 84.3	76.3 / 80.2	85.8 / 86.4	57.1 / 59.9
700	84.5 / 85.0	64.9 / 69.0	88.3 / 89.1	48.8 / 52.3
750	84.4 / 84.9	62.8 / 66.7	88.4 / 89.2	47.2 / 50.5
800	87.5 / 88.1	60.6 / 65.6	91.8 / 92.6	45.3 / 49.2
850	87.4 / 87.8	54.5 / 59.0	91.6 / 91.9	41.3 / 44.7
900	86.6 / 87.2	44.2 / 47.7	91.1 / 91.5	33.0 / 35.9
950	84.0 / 84.8	32.4 / 35.8	88.8 / 89.7	23.6 / 26.4
1 000	84.1 / 84.9	31.1 / 34.3	88.4 / 89.2	22.5 / 25.2

(b) Präzision und Vollständigkeit in Abhängigkeit von der Mindestzahl an Kollokationen bei einer Mindestbewertungszahl von 800

Mindestzahl der an Kollokationen	ohne SVM		mit SVM	
	Präzision	Vollständigkeit	Präzision	Vollständigkeit
1	87.4 / 88.1	60.6 / 65.6	91.8 / 92.6	45.3 / 49.2
2	90.7 / 91.6	49.6 / 52.9	93.6 / 94.8	36.2 / 38.9
3	90.3 / 91.4	39.9 / 41.2	93.4 / 94.7	30.6 / 32.5
4	89.8 / 90.9	37.8 / 39.5	93.0 / 94.3	28.7 / 30.4
5	90.0 / 91.3	33.8 / 35.3	93.3 / 94.8	26.0 / 27.7
6	90.1 / 91.5	31.9 / 33.1	93.8 / 95.5	24.1 / 25.3
7	90.4 / 92.0	30.6 / 31.6	94.5 / 96.4	23.3 / 24.6
8	90.7 / 92.4	28.9 / 29.7	94.4 / 96.3	22.5 / 23.7
9	91.1 / 92.9	27.3 / 27.9	94.0 / 96.0	20.9 / 21.9
10	90.7 / 92.6	26.3 / 26.7	94.9 / 97.2	20.1 / 21.0

die Zuordnung von Krankheitskonzepten zu Enzymklassen mit den gleichen stringenten Parametern, mit der die Krankheitsinformationen der BRENDA-Datenbank hinzugefügt wurden.

3.4.1 Übersicht der Zuordnungen von Krankheiten zu Enzymklassen

Tabelle 3.13 auf der nächsten Seite listet die Krankheitskonzepte und Enzymklassen mit den häufigsten Zuordnungen auf und gibt eine Übersicht der Verteilung von Krankheiten auf die sechs Hauptklassen der EC-Nomenklatur. Insbesondere den Hydrolasen sind bei annähernd gleicher Anzahl an Enzymklassen ungefähr doppelt so viele Krankheiten zugeordnet worden wie den Oxidoreduktasen und Transferasen; die Anzahl an Enzymklassen, denen mindestens eine Krankheit zugeordnet werden konnte, liegt ebenfalls mehr als doppelt so hoch.

Werden nur die Enzymklassen mit mindestens einer Krankheitszuordnung und die mittlere Anzahl an zugeordneten Krankheiten betrachtet (Abbildung 3.6), so zeigt sich eine ähnliche, wenn auch weniger deutliche Verteilung. Hier liegt die mittlere Anzahl bei dreizehn Zuordnungen je Enzymklasse und damit über dem Schnitt von zehn Zuordnungen je Enzymklasse bei den gewählten Parametern. Insbesondere die Hauptklassen mit vergleichsweise wenigen Enzymklassen wie die Ligasen liegen dagegen deutlich unter dem Mittelwert von zehn Zuordnungen. Jeweils fünf Enzymklassen und Krankheitskonzepte mit der höchsten Anzahl an Zuordnungen finden sich in Tabelle 3.14. Bei den Enzymklassen finden sich drei der Enzyme mit der höchsten Zahl von Nennungen innerhalb des Textkorpus wieder (siehe Tabelle 3.2 auf Seite 44). Bei den Krankheiten handelt es sich mit den arthritischen Erkrankungen, Diabetes Mellitus und einer Folgeerkrankung um besonders gut untersuchte Volkskrankheiten. Eine Liste mit den jeweils 50 Enzymklassen und Krankheitskonzepten mit der höchsten Zahl an Zuordnungen findet sich zusätzlich im Anhang A.4 auf Seite 129.

Die BRENDA-Datenbank hält Informationen zu den Gewebearten bereit, in denen die Enzymklassen identifiziert werden konnten. Eine Übersicht der am häufigsten vorgefundenen Annotation mit einem Gewebe der Enzymklassen mit mindestens einem zugeordneten Krankheitskonzept findet sich in Tabelle 3.15. Insgesamt enthielt die Annotation mehr als 300

Tabelle 3.13 · Verteilung der Krankheitszuordnungen auf die sechs EC-Hauptklassen. Die letzte Spalte gibt die mittlere Zahl von unterschiedlichen Krankheitskonzepten je Enzymklasse mit mindestens einem Krankheitskonzept an.

EC-Hauptklasse	Anzahl Enzymklassen	Anzahl Enzymklassen mit Krankheiten	Anzahl Krankheiten	Krankheiten je Enzymklasse
1 (Oxidoreduktasen)	1 026	107	511	4,78
2 (Transferasen)	1 067	91	386	4,24
3 (Hydrolasen)	1 168	247	991	4,01
4 (Lyasen)	363	47	215	4,56
5 (Isomerasen)	154	16	74	4,62
6 (Ligasen)	123	16	45	2,81

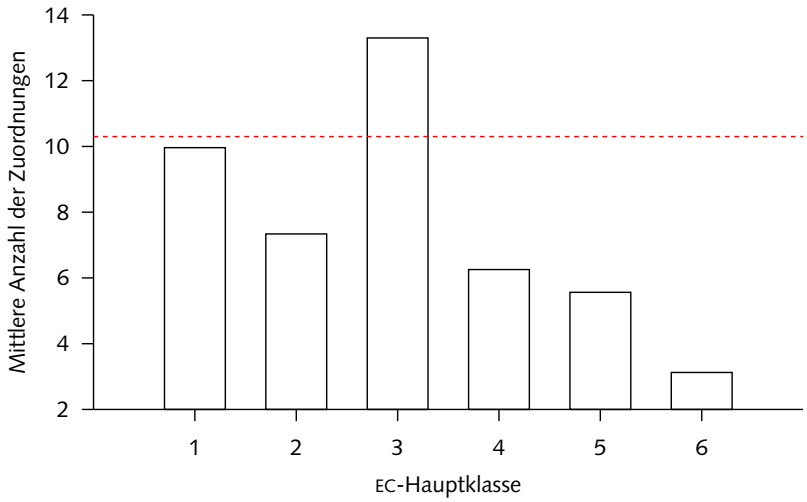


Abbildung 3.6 · Mittlere Anzahl der Krankheitszuordnungen je EC-Hauptklasse. Berücksichtigt wurden nur Enzymklassen mit mindestens einer Krankheitszuordnung. Die horizontale Linie markiert die mittlere Anzahl von Zuordnungen für alle Enzymklassen von 10,3 bei den gewählten Parametern.

Tabelle 3.14 · Liste der Enzymklassen und Krankheitskonzepte mit der größten Anzahl von Zuordnungen.

(a) Enzymklassen mit der höchsten Anzahl zugeordneter Krankheiten

Enzym	Anzahl Krankheiten
Peptidyl-Dipeptidase A (3.4.15.1)	121
Renin (3.4.23.15)	104
Pankreatische Elastase (3.4.21.36)	97
NO-Synthase (1.14.13.39)	92
Thrombin (3.4.21.5)	84

(b) Krankheiten mit der höchsten Anzahl zugeordneter Enzymklassen

Krankheit	Anzahl Enzymklassen
Rheumatoide Arthritis	100
Diabetes Mellitus, nicht insulinabhängig	77
Chronisches Nierenversagen	51
Diabetes Mellitus, insulinabhängig	49
Lupus Erythematosus	45

Gewebearten, woran unterschiedliche Zellkulturen einen deutlichen Anteil haben. In der Tabelle sind die zehn Gewebearten mit der höchsten Zahl an zugeordneten Krankheitskonzepten aufgeführt. Etwa ein Drittel der 524 Enzymklassen konnte in der Leber identifiziert werden; diesen Enzymklassen sind mehr als die Hälfte der Krankheitskonzepte zugeordnet. Einen ähnlich hohen Anteil haben die im Blut identifizierten Enzyme (77 in Erythrozyten und weitere 33 im Plasma).

3.4.2 Krankheiten und metabolische Pfade

Neben der Verteilung von Krankheiten auf die Hauptenzymklassen wurde die Verteilung auf die metabolischen Karten der KEGG-Datenbank überprüft. In der KEGG-Datenbank sind 95 metabolische Pfade verzeichnet, an deren Reaktionen im Menschen vorkommende Enzymklassen beteiligt sind. In 78 dieser Pfade konnten ein oder mehrere krankheitsrelevante Enzymklassen gefunden werden. Für eine erste Übersicht wurden 37 der KEGG-Karten in verschiedene Bereiche des Stoffwechsels unterteilt (siehe Anhang A.5 auf Seite 133); bei den restlichen Karten handelt es sich um spezialisierte Stoffwechselwege oder es finden sich nur wenige beteiligte humane Enzymklassen. Bei der Gruppierung zeigt sich eine im Verhältnis zu den beteiligten Enzymklassen gleichmäßige Verteilung der Krankheitskonzepte auf die sieben erstellten Bereiche (siehe Tabelle 3.16).

Tabelle 3.15 · Verteilung auf die Gewebearten der BRENDA-Datenbank für Enzymklassen mit mindestens einer Krankheitszuordnung. Die zweite Spalte zeigt die Anzahl an Enzymklassen mit einer Krankheitszuordnung, die in dem jeweiligen Gewebe gefunden wurden, die dritte Spalte die Gesamtzahl unterschiedlicher Krankheitskonzepte aller Enzymklassen mit dieser Gewebeannotation.

Gewebe	Anzahl Enzymklassen	Anzahl Krankheiten
Leber	164	844
Gehirn	79	628
Plazenta	82	623
Niere	82	623
Erythrozyten	77	490
Plasma	33	423
Bauchspeicheldrüse	33	391
Lunge	33	391
Hoden	22	385
Herz	36	361

Tabelle 3.16 · Verteilung der Enzymklassen mit zugeordneten Krankheitskonzepten auf verschiedene Bereiche des Stoffwechsels. Insgesamt 37 der metabolischen Karten der KEGG-Datenbank wurden in sieben Bereiche aufgeteilt (siehe Anhang A.5). Gezeigt ist die Anzahl der Karten eines Bereichs und die Zahl der in den Karten gelisteten Enzymklassen. Die dritte Spalte enthält den Anteil an Enzymklassen mit einer Krankheitsannotation und die Zahl unterschiedlicher Krankheitskonzepte für den jeweiligen Bereich.

Stoffwechsel	Anzahl Karten in KEGG	Anzahl Enzyme	Anzahl Enzyme mit Krankheiten	Anzahl Krankheiten insgesamt
Aminosäuren	17	361	81 (22 %)	393
Lipide	6	150	47 (31 %)	388
Zentralstoffwechsel	7	256	57 (22 %)	292
Nukleotide	3	161	39 (24 %)	278
Kofaktoren	9	171	34 (20 %)	223
Steroide	4	72	17 (23 %)	114
Fettsäuren	3	43	13 (30 %)	67

Tabelle 3.17 zeigt fünf Beispiele für einzelne KEGG-Karten mit einer im Vergleich zu den restlichen Pfaden hohen Anzahl krankheitsrelevanter Enzymklassen. Enzymklassen, die nur in dem jeweiligen metabolischen Pfad vorkommen – in der Tabelle als unikale Enzymklassen bezeichnet – geben an, inwieweit eine Überschneidung mit anderen Pfaden vorliegt. Ein hoher Anteil von Enzymklassen, die nur in einer KEGG-Karte vorkommen, erleichtert die Auswertung der Zuordnungen. So sind die 24 Enzymklassen des Sphingoglycolipd-Metabolismus im Durchschnitt nur in 1,3 KEGG-Karten zu finden, die der Glykolyse und Gluconeogenese dagegen in mehr als drei.

Ein Beispiel für die Darstellung der krankheitsrelevanten Enzymklassen in den einzelnen KEGG-Karten durch *tagKEGGMap* zeigt Abbildung 3.7 für den Inositolphosphat-Metabolismus, der eine besonders geringe Überschneidung mit anderen metabolischen Pfaden aufweist: 24 der 28 Enzymklassen sind in der KEGG-Datenbank nur diesem Pfad zugewiesen, fünf davon konnten insgesamt neun verschiedenen Krankheiten zugeordnet werden: Die Phospholipase C und die Phosphatidylinositol-Diacylglycerol-Lyase (EC-Nummern 3.1.4.3 und 4.6.1.13) kommen entgegen der KEGG-Annotation im Menschen vor und sind wie die Phosphoinositid Phospholipase C (EC 3.1.4.11) mit rheumatoider Arthritis und der Niemann-Pick-Krankheit assoziiert [22, 74]. Infektionen durch Staphylokokken stehen mit den Synonym ‚alpha-Toxin' der Phospholipase C in Verbindung [156]. Die in Pilzen vorkommende Phytase (EC-Nummer 3.1.3.26) wird als ein potentieller Auslöser von Asthma beschrieben [52]. Die Phosphoinositid 3-Kinase schließlich (EC-Nummer 2.7.1.137) ist Formen der Diabetes zugeordnet [88].

3.4.3 Vergleich der Zuordnungen von Krankheiten zu Enzymklassen mit Sequenzclustern

Für den Vergleich der den Krankheitskonzepten zugeordneten Enzymklassen mit einer Gruppierung von Enzymsequenzen aufgrund ihrer Sequenzähnlichkeit dienten die von Christian aus dem Spring bereitgestellten Daten (siehe Abschnitt 2.4.3 auf Seite 37). Die Daten enthielten 270 000 Cluster, davon bestanden 17 000 aus mehr als einer Sequenz. Hiervon wiederum beinhalteten 4 000 Cluster Enzymsequenzen, deren Annotation mit EC-Klassen sich in *Swiss-Prot* an einer der ersten drei Stellen un-

Tabelle 3.17 · Krankheiten und ihre Verteilung auf die metabolischen Karten der KEGG-Datenbank, sortiert nach der Anzahl der krankheitsassoziierten Enzyme. Der Anteil unikaler Enzymklassen beschreibt die Anzahl von Enzymklassen des Pfades, die an keinem anderen KEGG-Pfad beteiligt sind. Die letzten beiden Spalten zeigen die Anzahl der Enzymklassen des Pfades mit einer Zuordnung zu einem Krankheitskonzept und die Gesamtzahl an unterschiedlichen Krankheitskonzepten für diesen Pfade (in Klammern ist die mittlere Anzahl von Zuordnungen je Enzymklasse angegeben).

Bezeichnung des metabolischen Pfads in KEGG	Anzahl Enzyme	Anteil unikaler Enzymklassen	Anzahl Enzyme mit Krankheiten	Anzahl Krankheiten insgesamt
Glycerolipid metabolism	75	65 %	25	302 (12,1)
Purine metabolism	91	56 %	29	249 (8,6)
Arginine and proline metabolism	67	40 %	19	206 (10,8)
Glycolysis / Gluconeogenesis	36	31 %	20	185 (9,3)
Sphingoglycolipid metabolism	24	79 %	9	138 (15,3)

terschied. Der Vergleich der Enzymklassen dieser 4 000 Cluster mit den den Krankheitskonzepten zugeordneten Enzymklassen ergab in 209 Fällen eine Übereinstimmung – das heißt, mindestens zwei Enzymklassen des Sequenzclusters waren dem gleichen Krankheitskonzept zugeordnet. Einige Beispiele für Cluster und Krankheitskonzepte mit nicht direkt ersichtlichen und damit eventuell neuen Übereinstimmungen finden sich in Tabelle 3.18. Keine Berücksichtigung bei der Auswertung fanden größere Cluster, in denen eine der Enzymklassen nur durch eine Sequenz repräsentiert wurde oder solche mit zu allgemeinen Krankheitskonzepten wie ‚Herzversagen'. In der überwiegenden Zahl beruhten die Übereinstimmungen einerseits auf der Beteiligung von Enzymklassen wie den Cytochrom-c Oxidasen oder Proteinkinasen, die mit einer Vielzahl von Sequenzen in den Datenbanken vertreten sind. Andererseits führen multifunktionale Enzyme wie der *Swiss-Prot*-Eintrag Q23695 (*Bifunctional dihydrofolate reductase-thymidylate synthase*) mit mehr als einer anno-

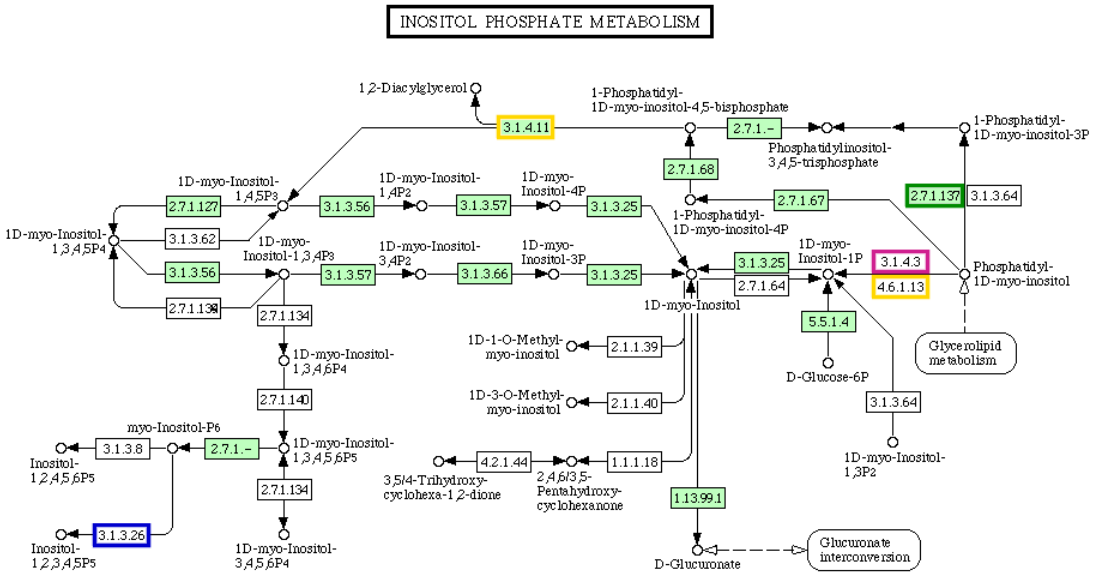


Abbildung 3.7 · Beispiel für die Darstellung einer metabolischen Karte der KEGG-Datenbank mit *tagKEGGMap*. Grün hinterlegte Felder kommen der KEGG-Annotation folgend im menschlichen Organismus vor. Enzymklassen mit ein bis fünf zugeordneten Krankheitskonzepten sind in den Farben blau, grün, gelb, orange und violett umrandet.

tierten Enzymklasse zu Clustern mit nur einer Enzymfamilie, aber mehreren EC-Nummern. In diesem Fall vereinigt das Protein die enzymatische Aktivität der Dihydrofolat Reduktase, EC 1.5.1.3, und der Thymidylat Synthase, EC 2.1.1.45.

3.4.4 Graphenanalyse

Struktur des Netzwerks

Für einen Vergleich der Netzwerke aus Enzymklassen oder Krankheiten mit anderen biologischen Netzwerken wurde die Topologie der Enzym- und Krankheitsgraphen untersucht, um so etwas über die Funktion der beteiligten Knoten zu lernen (siehe Abschnitt 2.4.3 auf Seite 38). Beide Netzwerke sind im Vergleich zu einem vollständig verbundenen Graphen spärlich konnektiert: der mittlere Knotengrad $\langle k \rangle$ (der Zahl an Kanten

Tabelle 3.18 · Vergleich der Enzymklassen von Krankheitskonzepten und Sequenzclustern. Die Anzahl der Cluster beschreibt die Anzahl unterschiedlicher Sequenzgruppierungen, die mindestens zwei dem Krankheitskonzept zugeordneten Enzymklassen enthielten. Die Anzahl der Sequenzen bezieht sich auf die Gesamtzahl aller in diesen Clustern gefundenen Sequenzen mit der entsprechenden EC-Nummer.

Krankheit	Anzahl Cluster	Enzymklasse	Anzahl Sequenzen
Cardiomyopathy	4	2.4.1.25 (4-α-Glucanotransferase)	13
		3.2.1.20 (α-Glucosidase)	128
		3.2.1.41 (Pullulanase)	60
Combined xanthine oxidase and aldehyde oxidase deficiency	11	1.1.1.204 (Xanthine dehydrogenase)	163
		1.1.3.22 (Xanthine oxidase)	50
		1.2.3.1 (Aldehyde oxidase)	136
Peripheral neuropathy	1	1.2.4.1 (Pyruvate dehydrogenase)	83
		2.2.1.1 (Transketolase)	24

eines Knoten) liegt deutlich unter dem maximal möglichen Knotengrad von $n - 1$, wobei n für die Knotenzahl steht.

Verglichen wurden die Netzwerkeigenschaften der Graphen mit einem zufälligen Netzwerk (*Random Network*, [56]) mit der gleichen Anzahl an Knoten und mittlerer Konnektivität (siehe Tabelle 3.19). In einem zufälligen Netzwerk folgt die Verteilung der Knotengrade einer Poisson-Verteilung. Zufallsnetzwerke mit einer großen Zahl von Knoten und mit $\langle k \rangle \ll n$ gehören zu den am schnellsten zu traversierenden Netzwerken, sie haben einen kleinen Durchmesser L [59]. Aufgrund der zufälligen Verteilung der Kanten erfolgt keine Gruppierung der Knoten, der Gruppierungskoeffizient C ist daher nahe null. Dies ist insbesondere für den Enzymgraphen nicht der Fall: der Gruppierungskoeffizient von 0,14 ist mehr als dreimal so groß wie der des gleich großen Zufallsgraphen. Sowohl Krankheits- als auch Enzymgraph entsprechen damit einem sogenannten *Small World*-Graphen[2], dessen Netzwerke als regelmäßige Gitter von Knoten beginnen, bei denen dann Kanten zufällig umverteilt werden und so Abkürzungen durch den Graphen schaffen. Diese Art von Graphen

[2] Die bekannten *six degrees of separation*, nach der jeder Mensch von jedem anderen Menschen in einem sozialen Netzwerk höchstens sechs Schritte entfernt ist, basieren auf einem *Small World*-Netzwerk [65].

hat aufgrund der benachbarten Knoten im Gitter einen deutlich höheren Gruppierungskoeffizienten C als Zufallsgraphen, aufgrund der Abkürzungen aber einem ähnlichem Durchmesser L [158].

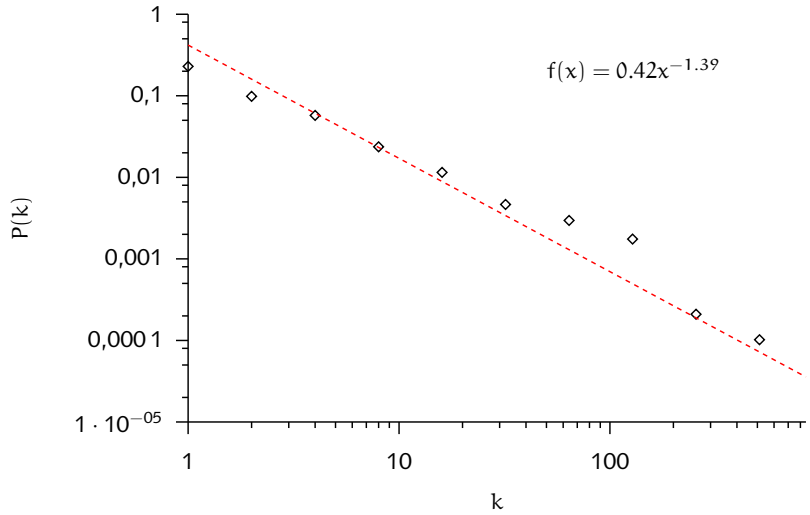
Sowohl in Zufallsnetzwerken als auch in *Small World*-Graphen kommen fast keine hochkonnektiven Knoten vor. Die Wahrscheinlichkeit P(k) für den Knotengrad k verringert sich exponentiell. Die Netzwerke sind homogen und weisen Knoten mit ähnlicher Konnektivität auf. Wie in Abbildung 3.8 zu sehen, folgt die Verteilung der Knotengrade von Enzym- und Krankheitsgraph dagegen einer Potenzfunktion. Damit entsprechen sie dem Modell des sogenannten *Scale-Free*-Graphen und weisen sehr unterschiedliche Knotengrade auf [13].

Subgraphen im Krankheitennetzwerk

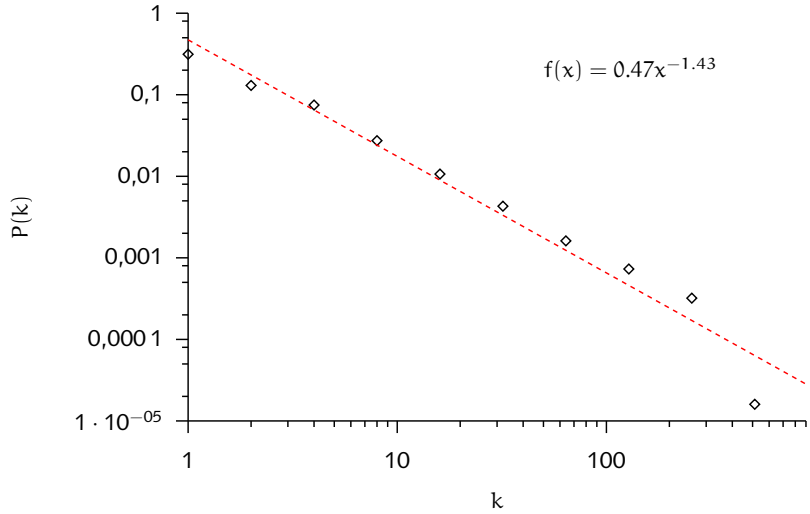
Bei der Suche nach nicht vollständig verknüpften Subgraphen (siehe Abschnitt 2.4.3 auf Seite 39) war eine Beschränkung auf Knoten mit weniger als 50 Kanten notwendig, um die Zahl der zu untersuchenden Krankheitskombinationen überschaubar zu halten. Mit dieser Begrenzung konnten 353 Subgraphen aus vier Krankheitsknoten gefunden werden, bei denen nur zwei Knoten nicht verknüpft waren. Abbildung 3.9 auf Seite 72 zeigt dazu ein Beispiel eines Subgraphen mit einer niedrigen mittleren Konnektivität von 8,5 Kanten je Knoten. Die Knoten der Krankheitskonzepte ‚familiäre Dystonie' und ‚Phenylalanin-Hydroxylase-Defizienz' sind nicht durch eine Kante miteinander verbunden, es gibt aber eine indirekte Verbindung über gemeinsame Enzymklassen mit den Krankheiten ‚Hyperphenylalaninaemie' und der ‚Dihydropteridin-Reduktase-Defizienz'. Die Zusammenhänge werden in Abschnitt 4.3.3 auf Seite 115 diskutiert.

3.5 Programm und Visualisierung

Das in dieser Arbeit entwickelte Programm umfasst zusammen mit den Testfunktionen 14 000 Zeilen und integriert viele Aspekte des *Text Minings*: den Aufbau eines Textkorpus, die Extraktion der im Korpus enthaltenen Informationen sowie die Visualisierung der Daten. Die modulare Implementierung erlaubt die Änderungen von Funktionalitäten (zum Beispiel Negationserkennung, automatische Zusammenfassung, etc.), ohne



(a) Verteilung der Knotengrade für den Graph aus Krankheitskonzepten



(b) Verteilung der Knotengrade für den Graph aus Enzymklassen

Abbildung 3.8 · Die Verteilung der Knotengrade für das Netzwerk aus Krankheitskonzepten (a) oder Enzymklassen (b). Aufgetragen ist die Wahrscheinlichkeit $P(k)$ gegen den Knotengrad k in einer doppelt logarithmischen Darstellung (nach Barabasi *et al.*, [14]). Die in den Graphen vorgefundenen Knotengrade wurden logarithmisch gruppiert. Das Bestimmtheitsmaß R^2 der Trendlinie lag bei Abbildung (a) bei 0,93 und in Abbildung (b) bei 0,96.

Tabelle 3.19 · Eigenschaften der Netzwerke aus Krankheitskonzepten oder Enzymklassen. Angegeben sind die Anzahl an Knoten n, die durchschnittliche Konnektivität $\langle k \rangle$, der Durchmesser des Netzwerks D sowie die durchschnittliche Pfadlänge L und der Gruppierungskoeffizient C. Beiden Netzwerken sind zum Vergleich die Eigenschaften eines zufälligen Graphen mit der gleichen Anzahl an Knoten und mittlerer Konnektivität gegenübergestellt.

Eigenschaft	Krankheitsgraph	Zufälliger Krankheitsgraph	Enzymgraph	Zufälliger Enzymgraph
n	1 406	–	524	–
$\langle k \rangle$	43,3	–	35,3	–
L	2,15	2,2	2,04	2,01
D	6	3	6	3
C	0,06	0,02	0,14	0,04

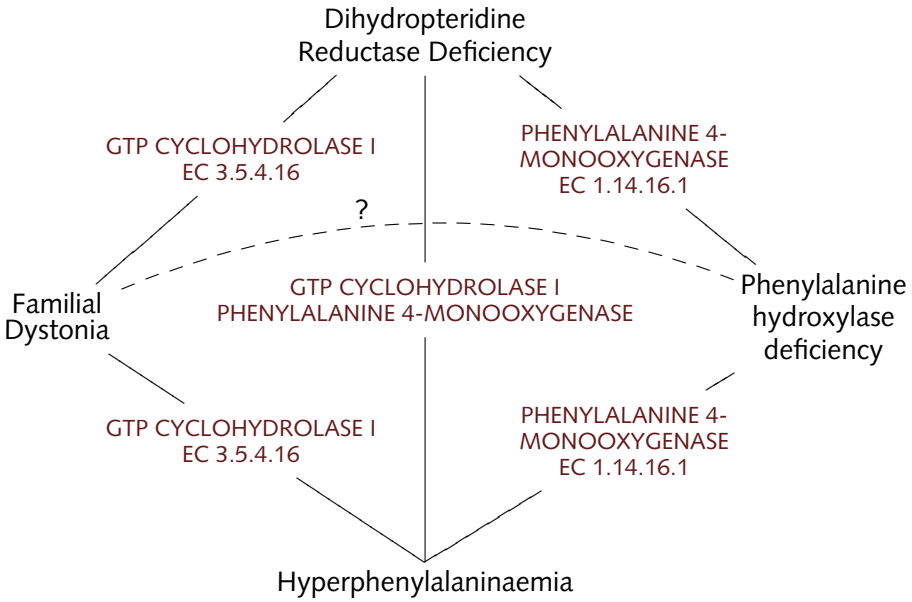


Abbildung 3.9 · Ein Beispiel für einen nicht vollständig verknüpften Subgraphen aus Krankheitskonzepten.

dabei andere Objektklassen und Methoden zu beeinflussen. Weitere Filter und Algorithmen sind damit leicht einzufügen. Eine Überprüfung der Konsistenz des Programms nach Veränderungen ist aufgrund der Implementierung von über 300 Funktionstests für alle Module und Funktionen jederzeit möglich. Die Konfiguration der verschiedenen Parameter und Pfade erfolgt über eine separate Steuerdatei.

Visualisierung großer Datenmengen

Neben der Erstellung von Daten ist die Repräsentation der gesammelten Informationen ein ebenso wichtiger Bestandteil des *Text Mining*, um dem Benutzer ein Werkzeug zur Analyse der Daten in die Hand zu geben. Aufgrund der erwähnten hohen Konnektivität einzelner Knoten des vorliegenden Netzwerks aus Enzymklassen und Krankheitskonzepten war es wichtig, eine Begrenzung auf relevante Informationen vornehmen zu können. Das zur Visualisierung verwendete Paket *Touchgraph* (siehe Abbildung 3.10 für ein Beispiel) bietet mehrere Vorteile. Es lassen sich ohne Verzögerung Teile des Graphen ein- oder ausblenden, was eine interaktive Auswertung der Datenmenge erleichtert. *Touchgraph* erlaubt eine Begrenzung der Anzeige auf die nähere Umgebung eines Knoten und somit eine Fokussierung auf ein für den Benutzer interessantes Gebiet.

Optional lassen sich zwischen den Krankheitskonzepten zusätzliche Kanten aufgrund der Ähnlichkeit der ihnen zugrunde liegenden Sätze einblenden. Dadurch ist es möglich, auch bei der Betrachtung kleinerer Subgraphen Krankheiten in die Analyse einzubeziehen, die nicht direkt über Enzyme, sondern wegen der Ähnlichkeit ihrer beschreibenden Texte miteinander in Beziehung stehen. Die für jeden Krankheitsknoten einblendbare Kurzzusammenfassung (siehe Abschnitt 2.4.2 auf Seite 36) vermittelt einen Überblick über die Krankheit. Diese sind mit den Dokumenten der *PubMed*-Datenbank verknüpft, aus denen sie extrahiert wurden, und erlauben damit eine schnelle Verifikation der Zuordnung sowie eine Einarbeitung in ein unbekanntes Themengebiet.

Gruppierung von Krankheiten

Das *TreeView*-Programm ermöglicht eine Betrachtungsweise des Netzwerks nach einer hierarchischen Gruppierung der Krankheiten aufgrund

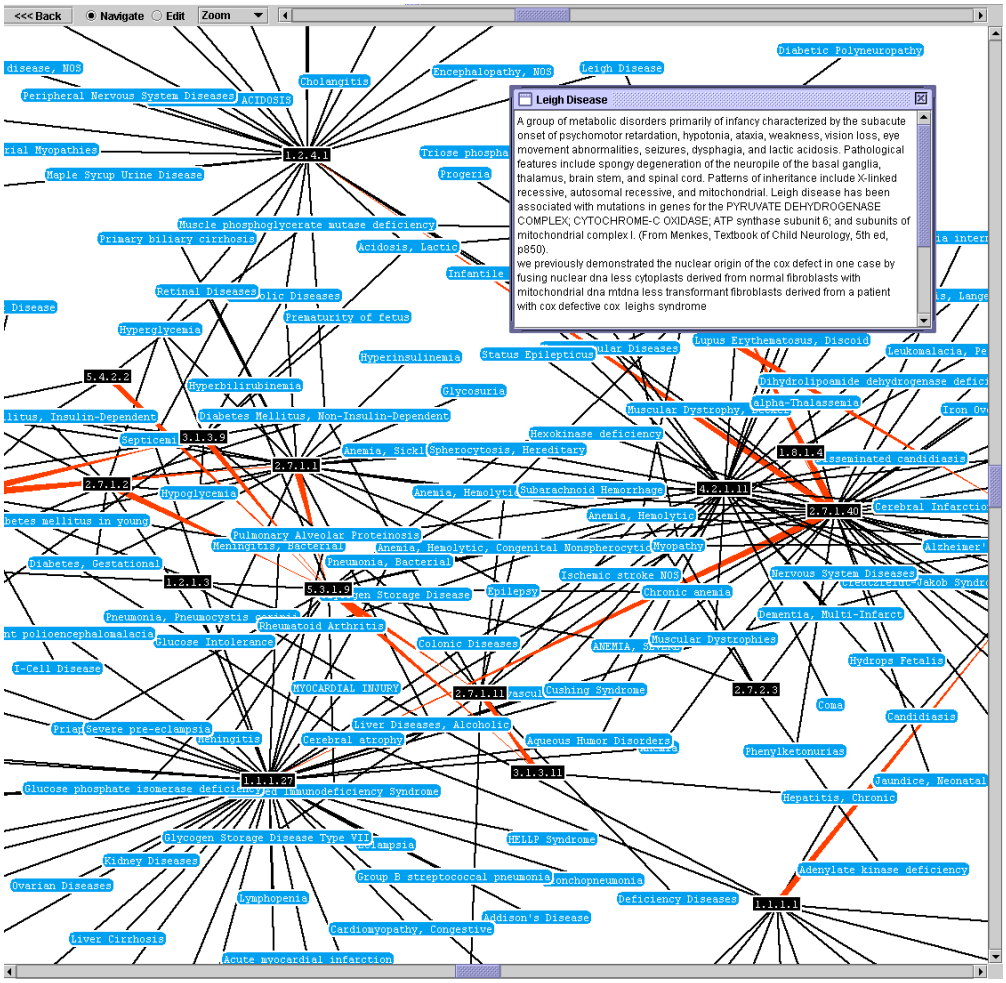
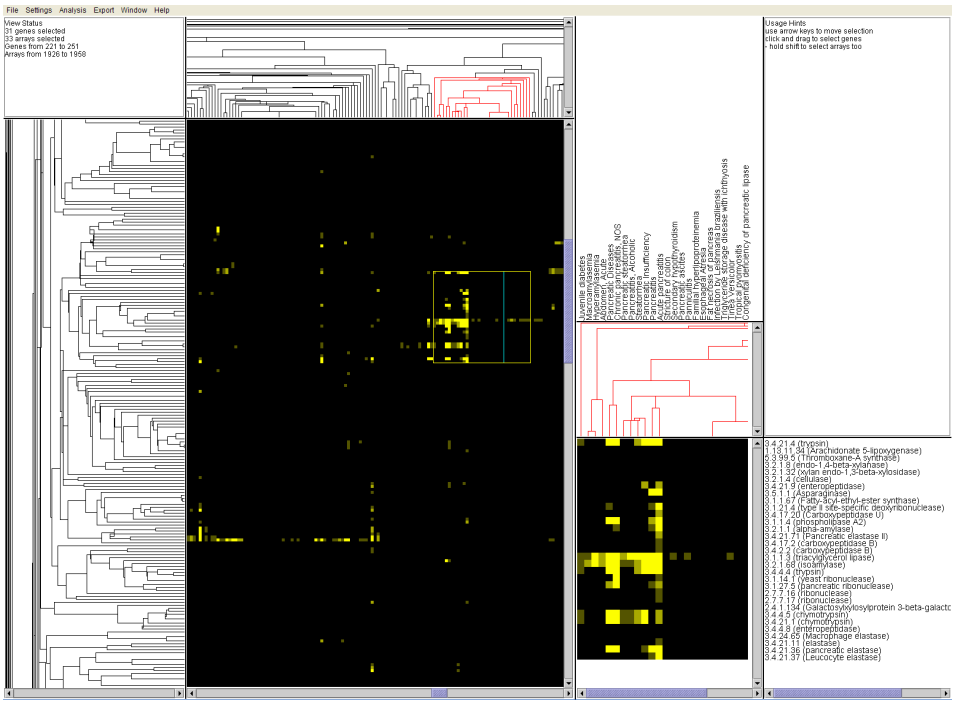
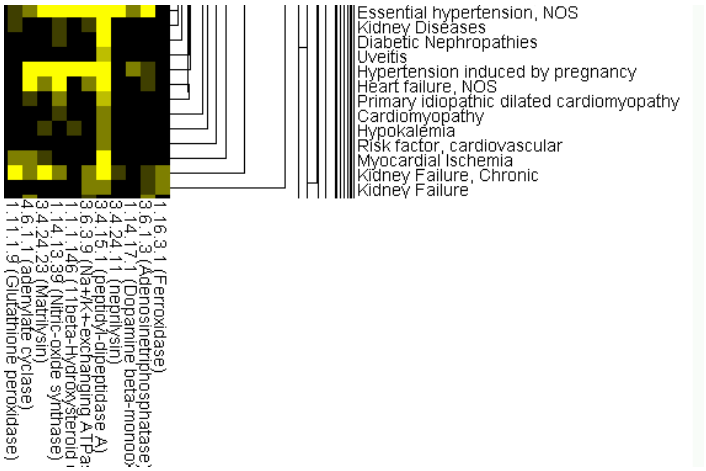


Abbildung 3.10 · Visualisierung des Netzwerks aus Enzymklassen und Krankheitskonzepten am Beispiel von Enzymklassen der Glykolyse und Gluconeogenese. Blaue Knoten repräsentieren Krankheiten (mit einer eingeblendeten Kurzzusammenfassung der Leigh-Krankheit), schwarze Knoten Enzymklassen. Schwarze Kanten stellen eine Zuordnung zwischen einer Enzymklasse und einer Krankheit dar, rote Kanten eine Substrat-Produkt-Kette zwischen zwei Enzymklassen.

gemeinsamer Enzymklassen. Im Gegensatz zu *Touchgraph* zeigt die Darstellung mit *TreeView* alle Zuordnungen gleichzeitig (Abbildung 3.11a). Die Gruppierung erlaubt die schnelle Identifikation von miteinander assoziierten Krankheiten oder Syndromen. Abbildung 3.11b zeigt als Beispiel für eine solche Gruppierung verschiedene mit Diabetes in Verbindung gebrachte Folgeerscheinungen. Neben bekannteren wie Nierenversagen oder Entzündungen des Auges finden sich auch weniger bekannte Erscheinungen wie die diabetische Cardiomyopathie [47, 142].



(a) Darstellung des Netzwerks mit *TreeView*



(b) Detailansicht

Abbildung 3.11 · Visualisierung des Netzwerks aus Enzymklassen und Krankheiten mit *TreeView*. Abbildung (a) zeigt die Darstellung nach der hierarchischen Gruppierung der Krankheiten aufgrund gemeinsamer Enzymklassen, Abbildung (b) eine Ausschnittsvergrößerung der mit Diabetes assoziierten Folgeerscheinungen.

4 DISKUSSION

> Scientists would rather share each other's underwear than use each other's nomenclature.
>
> *Keith Yamamoto*, UCSF

Die folgende Auswertung der Ergebnisse lässt sich in drei Abschnitte gliedern: Im ersten wird der Aufbau einer Dokumentensammlung sowie die Analyse mit linguistischen Methoden besprochen. Im zweiten Abschnitt folgt ein Vergleich der verschiedenen Zuordnungsmethoden, mit deren Hilfe Enzyme mit Krankheiten in Verbindung gebracht werden können. Der dritte Abschnitt behandelt die Präsentation der Daten und die mögliche Erstellung von Hypothesen, basierend auf den aus Krankheitsbegriffen und EC-Nummern generierten Netzwerken. Eine Zusammenfassung sowie ein Ausblick auf Erweiterungs- und Verbesserungsmöglichkeiten schließen das Kapitel ab.

4.1 Informationsextraktion & biomedizinische Publikationen

4.1.1 Aufbau eines Textkorpus

Der Großteil biologischer und medizinischer Erkenntnisse findet sich in wissenschaftlichen Publikationen wieder, welche in Zeitschriften, Konferenzberichten und in letzter Zeit auch zunehmend direkt im *World Wide Web* veröffentlicht werden. Die Anzahl der in dieser Arbeit durch Suchanfragen nach Enzymnamen und krankheitsrelevanten MESH-Begriffen gefundenen Artikel – insgesamt mehr als 100 000 Kurzzusammenfassungen mit zwanzig Millionen Worten – zeigt, dass im biomedizinischen Bereich ein ausreichend großer Textkorpus aus Einträgen der *PubMed*-Datenbank gebildet werden kann. Mit über zwölf Millionen Kurzzusammenfassungen ist sie eine der größten Literaturdatenbanken. Ihre mit MESH-Schlagwörtern annotierten und frei verfügbaren Einträge stellen zur Zeit den Standardtextkorpus dar, wenn es um die automatische Auswertung biologischer und medizinischer Informationen geht. Eine sich beschleunigende Wachstumsrate von 40 000 neuen Kurzzusammenfassungen im Monat [107] lässt darauf schließen, dass auch weiterhin für fast alle biologischen Fragestellungen ausreichende Textsammlungen zusammengestellt werden können.

Die aufgrund des Mangels an verfügbaren Volltexten notwendige Beschränkung auf Kurzzusammenfassungen hat auch einen Vorteil: sie besitzen die größte Informationsdichte. Während krankheitsbezogene Informationen in der Einleitung, den Ergebnissen und der Diskussion zu finden sind, ist dort das Verhältnis von relevanter Information zu für die Fragestellung irrelevanten Daten und Füllsätzen deutlich schlechter als in einer kompakten Kurzzusammenfassung [138].

Die in dieser Arbeit durchgeführte Verbindung von Informationsbeschaffung und Informationsextraktion erhöht die Informationsdichte weiter [143]: der Aufbau eines eigenen Textkorpus mittels geeigneter Suchbegriffe resultiert in einem höheren Anteil relevanter Kurzzusammenfassung, als er bei der Verwendung der gesamten *PubMed*-Datenbank zu erwarten wäre.

4.1.2 Identifikation von Enzymnamen

Zu den größten Problemen der Informationsextraktion in biologischen und medizinischen Texten gehört die automatische Namenserkennung – die Identifikation sogenannter *Entitäten*, also von Wörtern und Phrasen die ein ‚Objekt' beschreiben [35]. Dies gilt insbesondere für die Erkennung von Gennamen [102], RNA- und Proteinbezeichnungen [86], aber ebenso für die Namen chemischer Verbindungen [122]. Die Präzision automatisch agierender Namenserkennungssysteme liegt im Bereich von 75–80 %, weit von Erfolgsraten zwischen 90 % und 95 % entfernt, die in anderen linguistischen Domänen mittlerweile erreicht werden [48]. Daher erscheint es zunächst verwunderlich, dass in der Überprüfung einer Stichprobe von 100 *PubMed*-Dokumenten keine falschen und nur drei fehlenden Enzymzuordnungen gefunden werden konnten.

Das Enzymlexikon

Ein Grund für das gute Ergebnis bei der Erkennung von Enzymnamen stellt das mit Hilfe der BRENDA-Datenbank gebildete weitgehend vollständige Enzymlexikon dar. Selbst nach Entfernung von Abkürzungen mit weniger als vier Buchstaben und Zusammenfassung der unterschiedlichen Verwendung von Bindestrichen blieben im Durchschnitt noch mehr als sechs Namen je Enzymklasse übrig, ein Beleg für die hohe Zahl enthaltener Synonyme und Schreibweisen. Besonders deutlich wird dies bei den 180 Synonymen, die in der verwendeten Version der BRENDA-Datenbank für die Typ II-Restriktionsenzyme (EC 3.1.24.4) vorlagen (siehe Tabelle 3.1 auf Seite 43). Bei den Enzymen dieser Enzymklasse handelt es sich um Endonukleasen, die DNA sequenzspezifisch spalten.

Die Anzahl an Synonymen weist gleichzeitig auf die Problematik der variablen Namensgebung in der Biologie hin. Ein Grund für diese hohe Variabilität liegt darin, dass für die meisten biologischen Objekte entweder kein rigides Regelsystem zur Namensgebung existiert, oder ein solches durch eine Vielzahl von Ausnahmen geschwächt wird. Ein Gen wird wahlweise nach dem Phänotyp einer Mutation, der Ähnlichkeit zu anderen bekannten Genen, seinem Entdecker oder rein willkürlich benannt, selten jedoch nach seiner Funktion. Gene und Proteine mit bis zu 30 verschiedenen Namen sind keine Seltenheit [102]. Eine qualitativ hoch-

wertige Sammlung von Enzymnamen erlaubt die eindeutige Identifikation von Synonymen und legt die Verwendung eines Lexikon-gestützten Ansatzes nahe. Auch wenn Regelmäßigkeiten in Enzymnamen vorhanden sind, die sich durch statistische Verfahren auswerten lassen [162], kann zum Beispiel ein einfacher Schrägstrich zwei eigenständige Begriffe andeuten oder aber Teil des Namens eines einzigen Enzyms sein (*Ca^{2+}/calmodulin-dependent protein kinase*). In solchen Fällen führt der Vergleich mit einem Namenslexikon und das Prinzip der Zuordnung der Enzymklasse mit der größten Übereinstimmung zu einer eindeutigen Identifikation, während statistische Methoden Schwierigkeiten mit der Bestimmung der Namensgrenzen haben [39].

Die beschriebene Variabilität verstärkt sich im verwendeten Enzymlexikon durch die Gruppierung von Enzymnamen zu EC-Nummern aufgrund der katalysierten Reaktion noch weiter. So finden sich unter der EC-Nummer 2.7.1.37, den Proteinkinasen, eine Vielzahl von Enzymen in nur einer Gruppe zusammengefasst, die damit innerhalb des Textkorpus die meistgenannte Enzymklasse bildete (siehe Tabelle 3.2 auf Seite 44).

Enzyme, die wie die alkalische Phosphatase für Laborzwecke genutzt werden, fanden sich ebenfalls häufig und führten zu den in Tabelle 3.3 auf Seite 44 vorgefundenen Dokumenten mit mehr als zehn Enzymnennungen. Eine Spitzenposition nahm hierbei eine Veröffentlichung ein, die alle zur Fingerabdruckanalyse eines bakteriellen Genoms verwendeten Restriktionsenzyme in der Kurzzusammenfassung erwähnte – die aber mit 200 Einträgen zur EC-Nummer 3.1.21.4 zusammengefasst vorliegen. Eine andere wiederum nannte nicht weniger als zwanzig Enzymklassen, mit denen zu lysierende Tumorzellen vorbehandelt worden waren. Solche Veröffentlichungen mit mehr als ein oder zwei Enzymnennungen führten ebenso wie die in Bezug auf die biologische Funktion der Enzyme recht grobe Unterteilung in Enzymklassen zu der beobachteten hohen Konnektivität der Enzym- und Krankheitsnetzwerke.

Schwache Kontextanalyse

Während Synonyme über einen Thesaurus wie das verwendete Enzymlexikon zu handhaben sind, stellen für diese Arbeit die in biologischen Tex-

ten häufig vorkommenden Homonyme[1] ein größeres Problem dar. Viele Enzymnamen sind bestenfalls als unglücklich gewählt zu bezeichnen, vor allem wenn sie eigentlich schon eindeutig belegt sind. Steht das im Text gefundene Wort *Cat* für das Enzym *Catalase*, die nicht verwandte *Chloramphenicol transferase* – oder doch ganz schlicht für eine Katze? Ein anderes unschönes Beispiel ist die Abkürzung DNA für die *Nucleotidase*, EC 3.1.3.31.

Eine Disambiguierung des Wortsinns über seinen Kontext oder die Satzstruktur ist insbesondere in Kurzzusammenfassungen aufgrund der hohen Informationsdichte und verschachtelten Grammatik nur eingeschränkt möglich. Für die hier bearbeitete Aufgabenstellung erwiesen sich Homonyme jedoch als ein geringeres Problem. Nicht eindeutige Enzymnamen, die verschiedene EC-Nummern beschreiben, machen nach Ausschluss von Synonymen mit weniger als vier Buchstaben mit 2,1 % nur einen geringen Anteil des Enzymlexikons aus. In fast allen Fällen findet sich zudem in den Kurzzusammenfassungen neben der homonymen Abkürzung eines Enzymnamens noch ein anderer – meist vollständiger – Name, so dass aufgrund der angewendeten Kontextanalyse eine eindeutige Zuordnung möglich war. In der ausgewerteten Stichprobe konnten so elf Homonyme eindeutig einer EC-Nummer zugeordnet werden.

Da in dieser Arbeit die Zusammenhänge zwischen Krankheiten und Enzymen untersucht wurden, war eine Unterscheidung von Protein, RNA oder Gen zudem nicht notwendig. Eine Mutation des Promotors eines Enzymgens, die zu einer schwächeren Expression und in Folge dessen zu einem klinischen Phänotyp führt, ist genauso relevant wie ein Aminosäureaustausch im aktiven Zentrum des Enzyms mit den gleichen Folgen. Während für eine mechanistische Studie, bei der Details über die Enzymaktivität extrahiert werden sollen, eine solche Unterscheidung zwingend notwendig sein mag, sind sie bei der Zuordnung kausaler Zusammenhänge eher hinderlich und erschweren eine Verallgemeinerung der Assoziationen.

[1] Ein Synonym bezeichnet mehrere Worte, die das gleiche Konzept beschreiben. Homonyme dagegen sind Wörter, die für verschiedene Konzepte stehen. So kann zum Beispiel das Wort ‚Kiefer' im Deutschen sowohl einen Baum als auch den menschlichen Kiefer bezeichnen.

Fehlerquellen bei der Enzymidentifikation

Für die Vollständigkeit des aus der BRENDA-Datenbank erstellten Enzymlexikons spricht die Fehleranalyse in Abschnitt 3.1.1 auf Seite 45. In nur 10 % der nicht identifizierten Enzymnamen fehlte der Datenbank ein Synonym oder konnte in dieser nicht eindeutig zugeordnet werden. Unter den nicht gefundenen Enzymnamen waren auch solche, deren Namen selber schon ein Hinweis auf eine Krankheit enthalten:

— *. . . the kinase domain of anaplastic lymphoma kinase.*

Diese spezielle Kinase findet sich nicht in der BRENDA-Datenbank, die dazugehörige allgemeiner gehaltene MESH-Annotation *protein tyrosine kinase* dagegen durchaus.

In nur vier von fünfzig Fällen könnte eine Verbesserung der Identifikation durch eine Anpassung des Algorithmus unter der Beibehaltung eines Lexikon-gestützten Verfahrens erzielt werden. Um Verwechslungen mit MESH-Begriffen wie *topical* zu vermeiden, müssten dazu alle Namen aus dem Enzymlexikon entfernt werden, die in einem Lexikon der englischen Sprache zu finden sind. Eine Erkennung des Plurals eines Enzymnamens hätte in einem weiteren Fall geholfen.

In zwei Fällen konnte die Kontextanalyse keine eindeutige Zuordnung erreichen. So ist zum Beispiel das in der Phrase *cathepsin-kininogen complexes* gefundene Cathepsin Bestandteil von Enzymnamen neun verschiedener EC-Nummern (siehe Tabelle 4.1) und wurde in der Kurzzusammenfassung nicht genauer spezifiziert. Ohne weitere Informationen ist hier nur eine Zuordnung zu allen Enzymklassen möglich, was zu Lasten der Präzision geht.

Der größte Teil der nicht gefundenen Enzymnamen ließe sich nur über die Analyse des Volltextes identifizieren, welche aber weder verfügbar sind noch die hohe Informationsdichte der Kurzzusammenfassungen aufweisen. Die MESH-Annotation ist hier hilfreich, da sie sich auf den Volltext bezieht. Sie enthält aber keine Hinweise auf den Bezug der Enzymklassen zu krankheitsrelevanten MESH-Schlagwörtern.

Alternative Methoden zur Namenserkennung

Eine Eigenschaft Lexikon-gestützter Verfahren ist die Abhängigkeit von der Aktualität der Einträge: Es lassen sich prinzipiell nur Namen finden,

Tabelle 4.1 · Neun verschiedene Enzymklassen, deren Synonyme den Namen *Cathepsin* enthalten.

EC-Nummer	Empfohlener Enzymname	Synonym
3.4.11.1	Leucyl aminopeptidase	Cathepsin III
3.4.14.1	Dipeptidyl-peptidase I	Cathepsin C
3.4.16.5	Carboxypeptidase C	Cathepsin A
3.4.18.1	Cathepsin X	Cathepsin IV
3.4.22.1	Cathepsin B	Cathepsin B1
3.4.22.16	Cathepsin H	Cathepsin Ba, Cathepsin B3
3.4.22.38	Cathepsin K	Human osteoclast cathepsin K
3.4.22.43	Cathepsin V	Cathepsin L2
3.4.23.34	Cathepsin E	Cathepsin D-type proteinase

die bereits im Lexikon vorhanden sind, was angesichts des ständigen Wandels und Wachstums an biologischen Namen von Nachteil ist. So verzeichnet alleine die *Mouse Genome Database* zwischen 50 und 100 neue Namen oder Namensänderungen pro Woche [48]. Methoden, die ohne ein manuell annotiertes Lexikon auskommen, können dagegen auch potentielle neue Namen identifizieren, welche in den Standardlexika noch nicht zu finden sind. Diese Verfahren beruhen unter anderem auf der Analyse des zu annotierenden Text mittels statistischer und regelbasierter Methoden, zum Beispiel über einen Vergleich mit der im Lexikon verwendeten Orthographie, Morpheme oder n-Gramme[2] [162]. Synonyme werden als eigenständige Namen betrachtet [67] oder nur den semantischen Klassen DNA, RNA oder Protein zugeteilt [86, 93]. Für die vorliegende Arbeit kommen sie daher nicht in Betracht, zumal die höhere Vollständigkeit durch eine niedrigere Präzision erkauft wird.

Eine Erweiterung der Suche nach Enzymnamen durch einen Vergleich zwischen Einträgen des Lexikons und möglichen Kandidaten über die Editierdistanz oder den für Sequenzvergleiche entwickelten BLAST-Algorithmus [90] ist ebenfalls nicht sinnvoll. Der Vergleich der Enzymnamen untereinander (siehe Abschnitt 3.1.2 auf Seite 46) zeigt, dass keine klare Abgrenzung existiert, innerhalb der noch eine eindeutige Zuordnung zur richtigen EC-Nummer möglich wäre. Insgesamt 2 050 Namen lassen

[2] Abfolgen von n Termen

sich alleine mit nur einer Veränderung ineinander überführen. Ähnliche Versuche bei der Detektion von Proteinnamen durch eine Suche nach abweichenden Schreibweisen führten nur zu sehr geringen Verbesserungen [155].

Die im Vergleich zu statistischen Verfahren höhere Präzision des verwendeten Lexikon-gestützten Ansatzes stimmt mit der Zielsetzung der Arbeit überein, in jedem Stadium der Textprozessierung und Zuordnung von Enzymen zu Krankheiten eine möglichst geringe Zahl von falschen Zuordnungen vorzunehmen.

4.1.3 Textanalyse

Bestandteile von Texten

Eine automatische Auswertung von Texten, und insbesondere deren Vergleich, setzt eine geeignete Repräsentation der enthaltenen Informationen voraus. Jeder Autor hat seinen eigenen Schreibstil, verwendet unterschiedliche Worte oder bevorzugt einen anderen Satzbau. Wird in zwei Texten eine andere Reihenfolge der Themen verfolgt, so ist der Bezug zwischen beiden über einen einfachen Vergleich der Zeichenfolgen nicht mehr herzustellen.

Um diese Schwierigkeiten zu lösen, findet eine schrittweise Analyse von Dokumenten auf mehreren Ebenen statt [25]. Eine einfache Variante ist die Zerlegung eines Textes in seine Sätze und *Token*, elementare prozessierbare Einheiten. Die Bildung von *Token* aus von Leerzeichen begrenzten Zeichenfolgen ist bereits problematisch: Bindestriche können in einem vorhergehenden Schritt entfernt werden, spielen aber gerade bei chemischen und biologischen Namen oft eine wichtige Rolle und fassen Namen wie Glutamat-Ethylamin-Ligase zu einem Begriff zusammen. Die in dieser Arbeit angewandten Regeln (siehe 2.1.2 auf Seite 17) entfernen nur Gedankenstriche, also von Leerzeichen begrenzte Bindestriche. Das hat den Nachteil, dass verschiedene Schreibweisen des gleichen Namens zu unterschiedlichen *Token* führen. Ähnliche Schwierigkeiten bereitet der Schrägstrich, der oftmals nicht im Sinne eines ‚oder' verwendet wird: AML1/CBF steht für einen Komplex zweier Transkriptionsfaktoren und damit für ein eigenes biologisches Objekt.

In der vorliegenden Arbeit basierte nur die Identifikation von Enzymnamen auf der Textrepräsentation durch *Token*. Bei der Erkennung von Namen im Textkorpus stellte die *Token*-Bildung nur ein geringes Problem dar. Die Namen des Lexikons und die zu untersuchenden Texte wurden in der gleichen Weise behandelt. Dadurch konnten auch Enzymnamen, die während der Prozessierung in mehrere *Token* unterteilt wurden, noch im Text identifiziert werden. Vor einem Vergleich der Texte selbst ist es allerdings sinnvoll, die Anzahl zu vergleichender Elemente zu vereinheitlichen und auf die informativen Bestandteile zu reduzieren.

Die Bedeutung von Wörtern und Stammformen

Der Übergang von *Token* zu Wörtern ist fließend und beruht darauf, numerische Ausdrücke sowie Sonderzeichen zu entfernen. Da es in dieser Arbeit nicht um die Extraktion von Werten wie des Molekulargewichts oder der Thermostabilität von Enzymen ging, war der Informationsverlust akzeptabel.

Die Gewichtung von Worten ermöglicht die Ermittlung besserer Ähnlichkeitswerte von Texten über das *Vector Space Model*, als es mit einem einfachen Vergleich der verwendeten Wörter möglich wäre. Dieses Verfahren wird zum Beispiel von der *PubMed*-Datenbank zur Erstellung von Listen ähnlicher Dokumente verwendet [163]. Ebenso erlaubt es die automatische Annotation von Gensequenzen oder Proteinfamilien aufgrund der damit assoziierten Literatur [3, 4] oder die Themenfindung innerhalb einer wissenschaftlichen Domäne [76]. Die ermittelten Ähnlichkeiten zwischen Texten ermöglichten es in dieser Arbeit, während der Visualisierung der Enzym-Krankheiten-Netzwerke Krankheitskonzepte mit ähnlichen Beschreibungen gemeinsam zu gruppieren. Zusätzlich erlaubten sie die automatische Zusammenfassung der beschreibenden Texte aufgrund der durchschnittlichen Gewichte der einzelnen Sätze [54].

Dabei spiegelt die Gewichtung von Worten deren Bedeutung für einen Text wieder und basiert auf zwei einfachen Annahmen: Je öfter ein Wort in einem Text wiederholt wird, desto wichtiger ist es zur Beschreibung dieses Textes. Umgekehrt gilt, in je mehr Texten des Textkorpus ein Wort vorkommt, desto unwichtiger ist es für einzelne Texte. So beschäftigt sich ein Artikel, der sehr häufig das Wort ‚Diabetes' verwendet, vermutlich mit dieser Krankheit. Dagegen verliert die häufige Nennung eines ‚Patien-

ten‘ an Bedeutung, wenn sie in fast jedem Text des untersuchten Textkorpus vorkommt.

Aus der Gewichtung von Worten ergibt sich die Auswahl von Stoppwörtern, also der Worte, die aufgrund ihres Informationsgehaltes nicht zur Unterscheidung von Texten beitragen. Die in dieser Arbeit verwendete Liste reduzierte die Gesamtzahl an Worten des Textkorpus nach der Entfernung auf etwa 60 % (siehe Tabelle 3.6 auf Seite 50). Das so entstandene kleinere Vokabular ist auf die informationstragenden Elemente reduziert und erleichtert damit den Textvergleich.

Eine Rückführung von Wörtern auf ihre Stammformen erlaubt eine weitere Verbesserung der Vergleichbarkeit von Dokumenten. So führen nach der Umwandlung die Beschreibung ‚A wirkt aktivierend auf B‘ und ‚A aktiviert B‘ dazu, dass beide Texte die Stammform ‚aktiv‘ enthalten und damit einander ähnlicher werden. Durch die Entfernung von Stammformen, die nur in einem Text zu finden waren, ergab sich eine im Vergleich zur Verwendung von Worten noch bessere Reduktion des Gesamtvokabulars (siehe Tabelle 3.6 auf Seite 50). Dies betraf Worte beziehungsweise *Token*, für die es keine Stammform gibt: seltene chemische Formeln, Amino- und Nukleinsäuresequenzen sowie Eigen- und Autorennamen, die für einen Vergleich von Texten nicht informativ sind.

Konzepte und Ontologien

Eine weiterreichende Vereinheitlichung von Texten ließ sich durch die Verwendung von Konzepten erreichen. Für die Verwendung von Konzepten sprechen mehrere Vorteile:

- Unterschiedliche Repräsentationen des gleichen Konzepts, wie zum Beispiel ‚Bluthochdruck‘ und ‚Hypertonie‘, können durch eine Konzeptnummer repräsentiert werden, was eine Reduktion der Daten zur Folge hat.
- Die Zusammenlegung von Synonymen und Varianten auf ein Konzept verbessert, ähnlich wie die Rückführung von Worten auf ihre Stammformen, die Vergleichbarkeit von Texten.
- Zusammengesetzte Begriffe bleiben erhalten und sind damit von Texten zu unterscheiden, in denen die Worte einzeln vorkommen.

> So unterscheidet sich bei der Verwendung von Konzepten das Symptom *high blood pressure* von den drei eigenständigen Konzepten für die Worte *high*, *blood* und *pressure*. Im biomedizinischen Vokabular bestehen knapp 50 % der Konzepte aus solchen zusammengesetzten Begriffen, die besser als Beschreibungen aus einem Wort dazu geeignet scheinen, komplexe Sachverhalte zu erläutern [20].

Aus diesen Gründen erfolgten alle Textvergleiche sowie Kurzzusammenfassungen in dieser Arbeit auf der Basis von gewichteten Konzepten, auch wenn der Benutzer des entwickelten Programms optional Wort- beziehungsweise Stammformvergleiche verwenden kann.

Auf den ersten Blick ist es angesichts der Vorteile erstaunlich, dass so wenige Ansätze Konzepte zur automatischen Auswertung wissenschaftlicher Publikationen verwenden. Zwar gibt es Ausnahmen wie das SAPHIRE-System, dass Suchanfragen an die *PubMed*-Datenbank in eine Konzeptrepräsentation umwandelt [70], oder Simulationen des Swanson-Modells zur literaturgestützten Hypothesengenerierung, die Kombinationen von Konzepten untersuchen [99, 159] – die Mehrzahl der computerlinguistischen Ansätze nutzt allerdings nur Wörter und Wortkombinationen als Texteigenschaften [26]. Dies hat hauptsächlich zwei Ursachen: einen Mangel an frei verfügbaren biologischen Ontologien sowie die schwierige automatische Abbildung von Texten auf die in einer Ontologie enthaltenen Konzepte.

Konzepte und das Unified Medical Language System

Der Umfang der UMLS-Ontologie, ihr Schwerpunkt im biomedizinischen Bereich sowie die freie Verfügbarkeit gaben den Ausschlag für ihre Verwendung in dieser Arbeit. Insbesondere die im Laufe dieses Jahres abgeschlossene Integration des *Snomed*-Vokabulars – mit mehr als 100 000 Konzepten das Referenzvokabular der klinischen Terminologie – könnte zu einer weiteren Verbesserung der verwendeten Ontologie beitragen.

In der UMLS-Ausgabe von 2004 ist zudem bereits ein Großteil der GO-Ontologie integriert [17] und neuere oder noch nicht integrierte Konzepte lassen sich automatisch auf das UMLS-System abbilden [132]. Somit ist eine Erweiterung der Informationsextraktion auf biologische Themen wie die subzelluläre Lokalisation oder die Funktionsweise von Enzymen denkbar, ohne auf den Detailreichtum der GO-Ontologie verzichten zu müssen.

Für die Zuordnung von Textstellen zu Konzepten gibt es verschiedene Möglichkeiten. Die einfachste Methode ist die Suche nach direkten Übereinstimmungen zwischen dem prozessierten Text und Einträgen der Ontologie, sinnvoller ist aber die Verwendung eines Systems wie *MetaMap*. Das von der *Semantic Knowledge Representation*-Gruppe [141] entwickelte Programm verwendet mehrere linguistische Verfahren zur Abbildung von Texten auf Konzepte und ist bereits in vielen Anwendungen erfolgreich verwendet worden [5, 99, 121]. In Gegensatz zu anderen Programmpaketen [66, 96] ist *MetaMap* frei verfügbar und verwendet die Konzepte der UMLS-Ontologie für eine Zuordnung. Diese Kombination an Eigenschaften machte das Programm zur idealen Wahl für die hier bearbeitete Aufgabenstellung.

Verwendung von MetaMap

Die Abbildung des in dieser Arbeit verwendeten Textkorpus auf Konzepte des UMLS durch *MetaMap* führte gegenüber der Verwendung von Stammformen zu einer deutlichen Datenreduktion (siehe Tabelle 3.6 auf Seite 50). Die Reduktion des Gesamtvokabulars auf ein fünftel der Anzahl gefundener Stammformen ist erstaunlich hoch. Es stellt sich die Frage, wie zuverlässig die Konzepterkennung erfolgt und wie groß der Anteil der nicht von *MetaMap* identifizierten Konzepte ist.

Eine stichprobenartige Analyse von 50 Sätzen des verwendeten Textkorpus ergab eine Abdeckung von etwa zwei Drittel der im Text enthaltenen Phrasen. Bis zu 10 % der keinem Konzept zugeordneten Phrasen bestanden aus zusammengesetzten Wörtern und Artefakten wie ‚*12-year-old*‘ oder ‚3alpha‘.

Allerdings verlief die Zuordnung von Wörtern und Phrasen zu Konzepten nicht völlig fehlerfrei. Bei der manuellen Annotation von 1 000 Sätzen zeigte sich, dass das Hauptproblem für fehlerhafte Zuordnungen die Verwendung von Akronymen ist. So steht das Akronym RA sowohl für ‚rheumatische Arthritis‘ als auch für ‚Retinolsäure‘. Fehlzuordnung wie diese und andere ähnlich häufig auftretende Fehler konnten in einer Stopliste gesammelt und bei weiteren Durchgängen ausgeschlossen werden. Ebenfalls problematisch war die Zuordnung von Teilphrasen zu Konzepten durch *MetaMap*, obwohl für die vollständige Phrase eine genauere Beschreibung in der UMLS-Ontologie vorhanden ist. Der Begriff *hiv infec-*

tion wird zum Beispiel dem Konzept *opportunistic infection* zugeordnet, obwohl das Konzept *hiv infection* vorhanden ist. Der von *MetaMap* verwendete Algorithmus zur Bildung von *Token* bereitete zudem Schwierigkeiten bei der Prozessierung verschiedener Kombinationen von Sonderzeichen. Unter anderem führt eine mehrfache Interpunktion innerhalb von Sätzen zu Programmabbrüchen, was die Behandlung der für diese Arbeit wichtigen EC-Nummern erschwerte. Selbst nach einer Präprozessierung der Texte konnten knapp 2 % der Sätze des Textkorpus keinen Konzepten zugeordnet werden. Den Autoren des *MetaMap*-Programms sind die Fehler bekannt, mit entsprechenden Änderungen ist in einer der nächsten Versionen zu rechnen.

Dessen ungeachtet überwogen aber die Vorteile der Konzeptrepräsentation deutlich die beschrieben Schwächen. In einer Arbeit von Klein-van der Laaken *et al.* erreichte die Zuordnung von Konzepten durch *MetaMap* eine Präzision von 94 % bei einer zur Stichprobe ähnlichen Vollständigkeit von 64 % des gesamten Textes. Für die Ermittlung der Werte wurde ein von Experten manuell mit Konzepten annotierter Textkorpus verwendet [91]. Von den in der Studie nicht identifizierten Konzepten enthielten nur 7 % biomedizinisch relevante Informationen, was für die Qualität der Konzeptidentifikation durch *MetaMap* spricht. Auch bei der Zuordnung krankheitsrelevanter Konzepte in einer anderen Studie erreichte *MetaMap* vergleichbare Erfolgsraten [116].

4.2 Netzwerke von Krankheiten und Enzymklassen

4.2.1 Zuordnungen von Krankheiten zu Enzymklassen aufgrund gemeinsamer Nennungen

In Erweiterung der Informationsextraktion von Objekten versucht die Faktenextraktion Zusammenhänge zwischen diesen herzustellen. Dies erschwerte die Aufgabenstellung dieser Arbeit deutlich, da neben der korrekten Erkennung zweier Objekte noch eine Entscheidung getroffen werden muss, ob – und gegebenenfalls auch wie – diese beiden Objekte miteinander in Bezug stehen. Es gibt im Bereich der Computerlinguistik eine Vielzahl von Möglichkeiten, solche Zusammenhänge aus Texten zu extrahieren. Eine einfache und leicht an neue Aufgabenstellungen anzupassende Variante ist die Erstellung von Zusammenhängen zwischen zwei Begrif-

fen aufgrund ihrer Kollokationen (siehe Abschnitt 2.3 auf Seite 26). Der Ansatz beruht auf der Annahme, dass in Texten gemeinsam verwendete Begriffe häufig miteinander in einem Zusammenhang stehen. Verfahren dieser Art sind bereits erfolgreich für die Extraktion von Gen- [80, 143] und Proteininteraktionen verwendet worden [16, 100, 111].

Bei der vorliegenden Arbeit ging es um Beziehungen zwischen Enzymen und Krankheiten. Die durchgeführten Versuche zur Zuordnung von Enzymklassen zu Einträgen der OMIM-Datenbank zeigen, dass sich ein einfacher Zusammenhang bereits aufgrund einer entsprechend eindeutigen gemeinsamen Nennung herstellen lässt. Die bei der Verifikation mit den Annotationen der *Swiss-Prot*-Datenbank erhaltene niedrige Übereinstimmung erklärt sich dadurch, dass ein großer Anteil der in der Datenbank nicht gefundenen Zuordnungen richtig sind. Dies zeigt deutlich, dass eine automatische Annotation durch die Auswertung von Kollokationen möglich und sinnvoll ist, um Lücken in einer manuell annotierten Datenbank zu schließen.

Nicht gefundenen Zuordnungen basierten ähnlich wie in [80] beschrieben auf drei Faktoren: fehlende Namen im erstellten Enzymlexikon, Abweichungen in den Enzymnamen (wie die Eigenart von OMIM, das griechische α als @-Symbol darzustellen), und im Text nicht explizit erwähnte Zusammenhänge. Insbesondere Enzymkomplexe lassen sich ohne Nennung ihrer Bestandteile mit den verwendeten Methoden nicht identifizieren. So ist der OMIM-Eintrag 600212 (*Fatty Acid Synthase*; FASN) in *Swiss-Prot* neben der gefundenen EC-Klasse 2.3.1.85 noch sieben anderen Enzymklassen zugeordnet. Bei Bakterien und Pflanzen wird der Fettsäure-Synthase-Komplex noch von sieben verschiedenen Enzymen gebildet, die in Tieren zu einer langen, multifunktionellen Polypeptidkette fusioniert vorliegen. Die einzelnen Reaktionen sowie die sie katalysierenden Enzymklassen werden im OMIM-Eintrag nicht explizit genannt.

4.2.2 Zuordnungen über MeSH-Begriffe und Kurzzusammenfassungen

Eine Annotation von Enzymklassen über die in der OMIM-Datenbank vorhandenen Informationen hinaus benötigt eine automatische Erkennung von Krankheitsbegriffen. Dabei handelt es sich um ein weiteres Problem der Namens- oder Entitätenerkennung, für welches im Gegensatz zur Er-

kennung von Enzymnamen keine entsprechende Sammlung von Krankheitsbegriffen vorlag. Eine solche Quelle von Krankheitsbegriffen stellen die von der *PubMed*-Datenbank zur Annotation von Publikationen verwendeten MESH-Begriffe dar.

Die erreichte hohe Vollständigkeit von 88 % spricht für die Zuordnung von krankheitsrelevanten MESH-Begriffen zu Enzymklassen über ihrer Kollokation mit gefundenen Enzymnamen. Aufgrund der von den *PubMed*-Annotatoren erfolgten manuellen Annotation der Publikationen können auch Zuordnungen zwischen Begriffen hergestellt werden, die nicht explizit im Text der Kurzzusammenfassung erwähnt werden – solange sie von den Annotatoren als relevant betrachtet und mit MESH-Begriffen versehen wurden.

Dieser Vorteil der Annotation durch einen Experten ist zugleich mit Nachteilen verbunden. Die manuelle Annotation von Texten mit Schlagwörtern ist inhärent subjektiv (siehe Abschnitt 2.3.1 auf Seite 27). Zudem ist sie auf Einträge der *PubMed*-Datenbank limitiert. Eine Verwendung anderer Textdatenbanken ist mit diesem Verfahren nicht möglich. Ein zusätzlicher Nachteil liegt im Prinzip der ‚schwachen' Annotation: im Gegensatz zu einer vollständigen Annotation beschreiben MESH-Begriffe nur die *PubMed*-Dokumente, verweisen jedoch nicht auf die genaue Stelle im Text, die zu der jeweiligen Annotation führte. Dadurch kann ein Benutzer aufgrund der MESH-Annotationen erstellte Zuordnung von Enzymklassen zu Krankheitsbegriffen nicht direkt nachvollziehen. Eine Präsentation der Quelle der extrahierten Informationen ist nicht möglich und macht somit eine Einarbeitung in die Kurzzusammenfassungen, eventuell sogar in die vollständigen Publikationen notwendig. Gleichzeitig verhindert die ‚schwache' Annotation die Anwendung von linguistischen Methoden, welche die Satzstruktur oder das semantische Umfeld eines Krankheitsbegriffs auswerten könnten. Auch eine Distanzinformation zwischen MESH-Annotation und Enzymname existiert nicht. Die aufgrund dieser Probleme bei der Verwendung von MESH-Begriffen erreichte niedrige Präzision von 52 % ist nicht mit der Zielsetzung der Arbeit vereinbar, extrahierte Informationen direkt der BRENDA-Datenbank hinzufügen zu können.

Neben der geringen Präzision gibt es eine Reihe weiterer konzeptionelle Nachteile bei der Verwendung der MESH-Annotation. Nur wenige biologische Konzepte – wie zum Beispiel die Arten in einer Taxonomie – lassen

sich eindeutig in einer Baumstruktur wie der MESH-Hierarchie plazieren [146]. Für die meisten Definitionen reicht diese Datenstruktur nicht aus, was sich darin äußert, dass ein Großteil der Konzepte an mehreren Stellen in der Hierarchie vorkommen. Beispielsweise ist *Anisocoria*, eine Asymmetrie der Pupillen, unter den MESH-Bäumen *Nervous System Diseases*, *Eye Diseases* und *Pathological Signs and Symptoms* zu finden.

Ein weiterer Faktor ist die Zuordnung zu sehr unspezifischen Krankheitsbegriffen, entweder aufgrund einer zu allgemein gehaltenen Verwendung von MESH-Begriffen oder durch die Zusammenfassung verschiedener spezifischer Begriffe während der automatischen Zuordnung (siehe Abschnitt 2.3.1 auf Seite 26). So stehen Begriffe wie *brain diseases* oder *metabolic diseases* für eine Vielzahl von Krankheiten und bringen dadurch fast alle Enzymklassen miteinander in Verbindung. Eine Entfernung unspezifischer Krankheitsbezeichnungen aufgrund ihrer Position in der MESH-Hierarchie ist nicht möglich, da diese keine zuverlässigen Informationen über den Grad der Spezialisierung enthält – ein Problem, das auch von der GO-Ontologie bekannt ist [97]. Allgemeine Krankheitsbegriffe wie *hearing disorders* finden sich mit dem gleichen Abstand zur Wurzel der Hierarchie wie *Ochronosis* (Alkaptonurie), einer genau beschriebenen erblichen Stoffwechselerkrankung.

Neben den unspezifischen Krankheitsbeschreibungen trägt auch die hohe Zahl falscher Zuordnungen zu der sehr hohen Konnektivität des Netzwerkes aus Krankheiten und Enzymklassen bei. Ein Grund hierfür liegt in der Berücksichtigung von Kollokationen innerhalb des vollständigen *PubMed*-Dokuments. Werden in einer mit drei Krankheitsbegriffen indexierten Publikation zwei Enzyme genannt, so werden automatisch sechs Zuordnungen erstellt, von denen eventuell nur die Hälfte wirklich miteinander assoziiert sind. Dies resultiert in einer niedrigen Präzision von knapp 50 % und führt zu Netzwerken wie dem in Abbildung 4.1 auf Seite 94 gezeigten Beispiel für die an der Glykolyse und Gluconeogenese beteiligten Enzymklassen und deren Zuordnungen zu MESH-Begriffen. Eine Möglichkeit zur Verbesserung der Präzision böte die Limitierung auf *PubMed*-Dokumente mit nur genau einer Nennung einer Enzymklasse sowie einer Krankheit, was aber zu einem deutlichen Verlust an auswertbaren Kurzzusammenfassungen führt: Von den Dokumenten des erstellten Textkorpus enthielten mit 31 833 Dokumenten weniger als ein Drittel genau eine Nennung einer Enzymklassen sowie genau einen krankheitsrele-

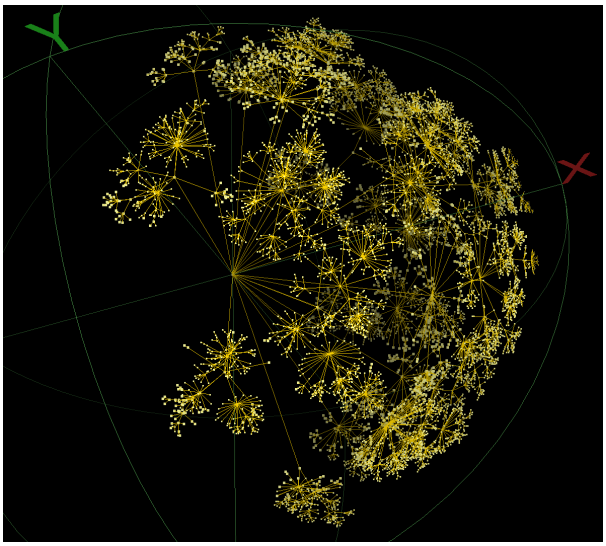
vanten MESH-Begriff. Eine Alternative zur Verwendung der MESH-Begriffe stellt die satzbasierte Zuordnung mittels krankheitsrelevanter Konzepte dar, mit der eine höhere Präzision und eine niedrigere Konnektivität erreichbar war.

4.2.3 Zuordnungen über Sätze und Konzepte

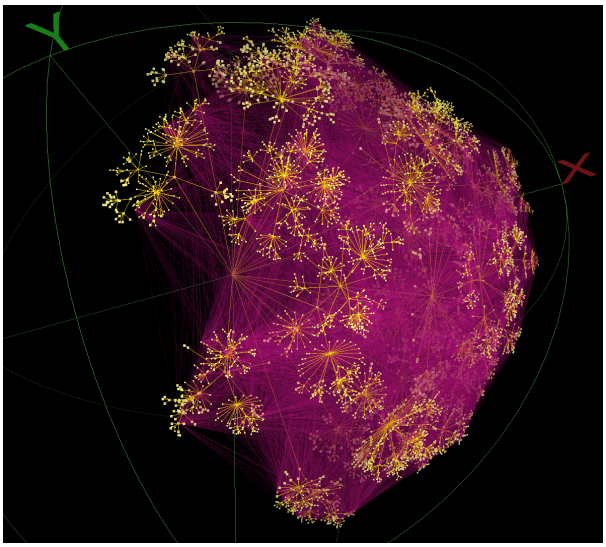
Die Größe eines auf Kollokationen untersuchten Textabschnitts hat einen direkten Einfluss auf die Präzision der daraus resultierenden Zuordnung zwischen den einzelnen Objekten. In einem Vergleich von Präzision und Vollständigkeit bei der manuellen Annotation biologischer Interaktionen aufgrund gemeinsamer Nennungen in Textbausteinen verschiedener Größe – Kurzzusammenfassungen, Sätzen und Phrasen – konnte gezeigt werden, dass sich ein satzbasierter Ansatz am besten eignet. Die von Ding *et al.* beobachtete Präzision von 64 % und eine Vollständigkeit von 85 % bei der Verwendung von Sätzen basierte auf der manuellen Auswertung von 300 Kurzzusammenfassungen aus insgesamt zehn Suchanfragen mit zwei biologischen Begriffen. Die Werte stellen einen Maßstab für das in dieser Arbeit vorgestellte Verfahren dar [50]. Auch andere Arbeiten unterstützen diese Beobachtung und konzentrieren die Auswertung auf kleine Textabschnitte [58, 109, 122].

Mittels der den Konzepten zugewiesenen semantischen Felder war die Identifikation von krankheitsrelevanten Konzepten leicht möglich (siehe Abschnitt 2.1.2 auf Seite 19). Bei der Auswertung der Zuordnungen von Enzymklassen zu Krankheitskonzepten (siehe Abschnitt 2.3.4 auf Seite 32) gibt es zwei unterschiedliche Betrachtungsweisen:

- *Extraktion der einzelnen Nennungen:* Bei der Bewertung der Extraktion jeder einzelnen Kollokation mit einem kausalen Zusammenhang zwischen Enzymklasse und Krankheit konnte eine Präzision und Vollständigkeit von 81 % erreicht werden (siehe Tabelle 3.9 auf Seite 55). Die erreichte Präzision übertrifft die Beobachtungen von Ding *et al.* bei vergleichbarer Vollständigkeit deutlich. Dies könnte an der anderen Aufgabenstellung liegen – die von Ding *et al.* durchgeführte Studie beschäftigte sich hauptsächlich mit den Zusammenhängen von biologischen Molekülen und Proteinen.



(a) krankheitsrelevante Begriffe der MESH-Hierarchie



(b) mit zusätzlichen Kanten zwischen MESH-Begriffen mit einer gemeinsamen Enzymklasse

Abbildung 4.1 · Abbildung (a) zeigt eine Übersicht des krankheitsrelevanten Baums der MESH-Hierarchie mit dem Programm *Walrus*. Abbildung (b) enthält zusätzliche Kanten zwischen zwei Krankheitsbegriffen, die mindestens eine an der Glykolyse oder Gluconeogenese beteiligte Enzymklasse aufgrund der automatischen Zuordnung gemeinsam haben.

- *Extraktion der Zuordnungen:* Eine höhere Präzision lässt sich unter Vernachlässigung des einzelnen Extraktionsergebnisses erzielen. Für eine Erweiterung von Datenbanken mit automatisch extrahierten Zusammenhängen spielt die Extraktion jeder einzelnen Nennung eines Zusammenhangs eine untergeordnete Rolle, solange dieser in mindestens einer anderen Publikation gefunden werden kann. Bei der Frage, *ob* eine Zuordnung gefunden wurde – statt der Frage, *wie oft* sie entdeckt wurde – verbesserte sich die Vollständigkeit bei gleichbleibender Präzision auf 84 %.

Ein Vergleich dieser Ergebnisse mit der relevanten Literatur fällt schwer. Vergleichbare Arbeiten, die ebenfalls mit Konzepten der UMLS-Ontologie arbeiten, enthalten entweder keine Angaben zu Vollständigkeit und Präzision des Verfahrens [122], verwenden andere Kennzahlen wie den Mittelwert von Präzision und Vollständigkeit [58] oder beschäftigen sich mit anderen Informationen: Für die Erkennung der Interaktionen biologischer Moleküle wurde eine Präzision von 73 % bei einer Vollständigkeit von 53 % beschrieben [125]. Wren und Garner gehen bei einer gemeinsamen Nennung von Genen und Kategorien der *Gene Ontology* innerhalb eines Satzes von einer Wahrscheinlichkeit von 83 % für einen Zusammenhang zwischen beiden Begriffen aus.

In der vorliegenden Arbeit beruhten fehlende Zuordnungen zwischen Enzymklassen und Krankheitskonzepten im Textkorpus von 1 500 Sätzen auf Fehlern bei der Prozessierung durch *MetaMap*, fehlerhaften Schreibweisen von Namen und der Unvollständigkeit des Enzymlexikons oder der UMLS-Ontologie. Falsch erstellte Zuordnungen lassen sich dagegen durch eine Vielzahl von Faktoren begründen. So können neben der Gegenüberstellung verschiedener Krankheiten innerhalb eines Satzes Negationen zu falschen Zuordnungen führen:

> — . . . *gene polymorphism for* angiotensin converting enzyme *is not linked to the development of* microalbuminuria *or established* diabetic nephropathy

Ebenso problematisch sind Sätze, die sich mit Labormethoden beschäftigen, zum Beispiel bei der Beschreibung von Isolationsmethoden:

> — *. . . however purified fibrin monomers isolated from plasma using both* reptilase *and* thrombin *exhibited delayed polymerization rates and the occurrence of* acquired dysfibrinogenaemia in liver disease *is therefore confirmed.*

Während es kleinere Textbausteine erleichtern, die richtigen Zusammenhänge aus einem Text zu extrahieren, geht damit ein Verlust an Vollständigkeit einher. Auf mehrere Sätze verteilte Erkenntnisse sind mit Methoden, die nur Kollokationen innerhalb eines Satzes berücksichtigen, nicht zu identifizieren:

> — *We evaluated the effect of enzyme-replacement therapy with recombinant human alpha-l-iduronidase in patients with* this disorder.

Ohne eine Diskursanalyse der vorhergehenden Sätze bleibt unklar für welche Krankheit eine Enzymtherapie gedacht ist. Dennoch ist die Fokussierung auf eine höhere Präzision, auch zu Lasten der Vollständigkeit, bei der Extraktion von Zusammenhängen aus großen Literaturdatenbanken wie der *PubMed* eine sinnvolle Strategie, da die in einem Textabschnitt nicht gefundenen Interaktionen aufgrund weiterer Kollokationen noch extrahiert werden können.

Filter zur Verbesserung der Präzision

Die erreichte Präzision bei der Verwendung von Krankheitskonzepten und Sätzen konnte durch die Anwendung unterschiedlicher Filter noch weiter verbessert werden. Direkt auswertbar war die von *MetaMap* vergebene Bewertungszahl der Konzeptzuordnung. Je höher die Bewertungszahl – und damit die Konfidenz in die Identifikation – eines Konzepts ist, desto geringer ist die Wahrscheinlichkeit einer falschen Zuordnung. Dies bestätigte sich bei der manuellen Durchsicht einer Auswahl von Konzeptzuordnungen. Insbesondere Konzeptbezeichnungen, die nur teilweise mit dem ursprünglichen Text übereinstimmen, oder für die mehrere Modifikationen notwendig sind, erhalten eine niedrige Bewertungszahl (siehe Tabelle 4.2, Beispiele 1 und 2). Eine Filterung nach dem Grad der Übereinstimmung hilft jedoch nicht bei Konzepten, bei denen das in der UMLS-Ontologie gewählte semantische Feld nicht den Erwartungen entspricht:

eine erhöhte Aktivität alleine ist keine Krankheit (Beispiel 3). Auch zu allgemein gewählte Konzepte wie *disease* in Beispiel 4 führen zu Schwierigkeiten.

Durch die Entfernung von Konzepten unter einer Bewertungszahl von 800 konnte die Präzision der Zuordnungen um 3 % gesteigert werden (siehe Tabelle 3.9(a) auf Seite 55). Eine Entfernung zu allgemein gehaltener Konzepte über Verwandschaftsbeziehungen ist, wie schon bei der Zuordnung zur MESH-Hierarchie, nicht möglich. Das semantische Netzwerk der UMLS-Ontologie ist nicht vollständig. Nicht allen Konzepten sind übergeordnete Begriffe zugewiesen, und die Definitionen können aufgrund unterschiedlicher Quellen inkonsistent sein [96, 114]. Das verhindert eine zuverlässige Filterung aufgrund der Verwandschaftsbeziehungen. Da die UMLS-Ontologie aus einer Vielzahl verschieden detaillierter Lexika aufgebaut wurde, besitzen die enthaltenen Konzepte je nach Herkunft eine unterschiedliche Spezifität. Daher ist der Abstand eines Konzepts zu einem allgemeineren Konzept kein geeignetes Ausschlusskriterium.

Durch den Einsatz eines zweiten Filters – der Mindestzahl an Kollokationen eines Krankheitsbegriffs und einer Enzymklasse – konnte die Präzision um weitere 3–6 % gesteigert werden (siehe Tabelle 3.9(b)). Bei der Bestimmung der Mindestzahl ist ein Kompromiss zwischen falschen Zuordnungen und richtigen, aber selten beschriebenen Zusammenhängen zu finden. Insbesondere höhere Mindestzahlen wirken sich deutlich auf die Gesamtzahl gefundener Zuordnungen aus. Für den verwendeten Textkorpus ist eine Entfernung aller nur einmalig gefundenen Kollokationen ausreichend, um einen Großteil des Präzisionsgewinns durch diesen Filter zu erreichen. Bei höheren Mindestwerten für die Kollokation werden richtige wie falsche Zuordnungen gleichermaßen entfernt, so dass sich die Präzision nur noch langsam verbessert. Bei falschen Zuordnungen mit mehr als drei gemeinsamen Nennungen im Textkorpus handelt es sich meist um systematische Fehler, wie zum Beispiel Doppeldeutigkeiten in der Ontologie:

> — *. . . three bk virus mutants forming clear large plaques like those of wt 501 but capable of transforming* rat *cells were derived from the recombinant virus carrying the* hind iii *c segment.*

Tabelle 4.2 · Beispiele für die Filterung von satzbasierten Zuordnungen über die von *MetaMap* ermittelten Bewertungszahl bei der Konzeptidentifikation. Sätze 1 und 2 zeigen niedrige Bewertungszahlen aufgrund einer nur geringen Übereinstimmung mit den Konzeptrepräsentationen. Satz 3 ist ein Beispiel für ein unerwartetes Konzept mit dem semantischen Feld *disease or syndrome*. Ein für die verfolgte Aufgabenstellung zu allgemeines Konzept findet sich in Satz 4.

Nummer	Satz	Konzept und Bewertung
1	Immunohistochemical staining of *lung tissue* with anti human neutrophil elastase. . .	Lung diseases (694)
2	Three bk virus mutants forming clear large *plaques* like those. . .	dental plaques (670)
3	The patients had normal or *increased activities* of. . .	increased activities (900)
4	. . . in inactive *tb* patients reveals the quiescent stage of the *disease*	Disease (1000), Tuberculosis (660)

Hier wird der Begriff RAT als Abkürzung für *recurrent acute tonsilitis* identifiziert, einer durch Streptokokken hervorgerufene Entzündung. Das hat zur Folge, dass sämtliche in Zusammenhang mit Versuchen an Ratten erwähnten Enzymklassen mit dieser Krankheit in Verbindung gebracht werden. Fehler dieser Art lassen sich durch eine höhere Mindestzahl an Kollokationen nicht vermeiden. Es besteht zwar die Möglichkeit, Akronyme bei der Konzeptzuordnung durch *MetaMap* auszuschließen, dies hilft allerdings nicht, wenn das Akronym wie in diesem Beispiel auch als vollständiges Wort in der Ontologie existiert. Eine separate Behandlung von Worten in Großbuchstaben ist mit *MetaMap* nicht möglich. Abhilfe könnte hier eine Präprozessierung der Texte machen, während der Akronyme und Abkürzungen durch ihre vollständige Schreibweise ersetzt werden [124].

Als dritter verwendeter Filter erzielte die Entfernung von Sätzen mit einer möglichen Negation eine nochmalige Verbesserung der Präzision um etwa 3 % (siehe Tabelle 3.9 auf Seite 55, ohne Negationen). Bei der

Durchsicht der annotierten Sätze zeigte sich, dass Verneinungen in nur knapp 3 % der untersuchten Sätze vorkommen. Dabei sind nicht alle Negationen automatisch falsche Zuordnungen. So kann zum Beispiel die beschriebene fehlende Aktivität eines Enzyms (*no activity of*) ursächlich für eine metabolische Krankheit sein, anstatt ein negatives Ergebnis zu kennzeichnen:

> — *There was no* alpha-d-mannosidase *activity in the hair roots of the patient with* mannosidosis.

Insbesondere bei einer höheren Mindestzahl an Kollokationen wird die Verbesserung der Präzision durch die Entfernung von Negationen geringer. Viele der aufgrund von Verneinungen hergestellten Fehlzuordnungen finden im Textkorpus selten Erwähnung und werden durch den Kollokationsfilter entfernt. Bei weniger stringenten Einstellungen kann sich die Entfernung von Sätzen mit einer Negation lohnen, um die Präzision zu steigern (siehe Abbildung 3.3 auf Seite 56).

Erstaunlicherweise lässt sich bei der Begrenzung auf Sätze, in denen genau eine Enzymklasse und eine Krankheit vorkommen, fast keine Veränderung der Präzision feststellen (siehe Tabelle 3.10 auf Seite 57). Dabei können insbesondere bei Gegenüberstellungen und Vergleichen von Krankheiten in einem Satz sowohl richtige als auch falsche Zuordnungen vorkommen:

> — *. . . the granulomatous inflammation in* Crohn's disease *differs from that of* sarcoidosis *in which there is a striking elevation of* angiotensin converting enzyme *in serum.*

Der Anteil richtiger Zuordnungen steigt jedoch bei der Verwendung von Sätzen mit mehreren Nennungen im gleichen Masse wie die der falschen. Die Begrenzung auf eindeutige Sätze ist daher nicht sinnvoll. Auch bei der Verwendung aller Sätze lässt sich die Präzision durch die Anwendung moderater Filterkriterien (Bewertungszahl 800, mindestens zwei Kollokationen, Entfernung von Negationen) ohne drastische Einbußen in der Vollständigkeit von 82 % auf über 91 % steigern (siehe Abbildung 3.4 auf Seite 58).

Hinweise aus dem Kontext

Bei der manuellen Annotation von Texten zur Bestimmung der Präzision und Vollständigkeit zeigte sich, dass eine grobe Unterteilung der Sätze in verschiedene Klassen möglich ist. Sätze, die den genetischen Hintergrund einer Krankheit beschreiben, unterscheiden sich oft deutlich von solchen, die deren klinischen Aspekte zum Inhalt haben. Beschreibungen von Labormethoden zur Analyse von Proteinen oder zellulären Strukturen verwenden eine andere Wortwahl, oder anders formuliert: Die gefundenen Enzymklassen und Krankheitskonzepte treten in unterschiedlichem Kontext auf. In der Computerlinguistik gibt es verschiedene Ansätze zur Disambiguierung mehrdeutiger Wörter aufgrund des Kontexts, in dem sie verwendet werden. Dabei finden meist Methoden aus dem maschinengestützten Lernen Anwendung, um anhand eines Trainingsdatensatzes die unterschiedlichen Bedeutungen eines Wortes über das unmittelbare Umfeld klassifizieren zu können [29]. Insofern lag es nahe, die Methoden zur Disambiguierung von einzelnen Wörtern auf die Klassifikation von Sätzen anzuwenden – handelt es sich um einen Satz, in dem ein Zusammenhang zwischen einem Enzym und einer Krankheit hergestellt wird?

In der Biologie sind ähnliche Verfahren bereits verwendet worden. So benutzt Craven hierarchische *Hidden Markov*-Modelle zur Unterscheidung von relevanten und irrelevanten Sätzen bei der Extraktion von Informationen zur subzellulären Proteinlokalisation [43]. In Arbeiten von Craven wird ein *Naive Bayes*-Verfahren zur Klassifikation von Sätzen anhand ihrer Stammformen angewendet [41, 42].

Das Fehlen eines sinnvoll annotierten Trainingsdatensatzes machte die Erstellung eines eigenen Datensatzes aus 1 000 manuell annotierten Sätzen notwendig. Zwar steht der biologischen Computerlinguistik mit dem *Genia*-Korpus seit kurzem eine Sammlung von 2 000 manuell annotierten Kurzzusammenfassungen mit fast 100 000 Annotation zur Verfügung [87]. Dieser eignete sich aber nicht als Trainingsdatensatz für die Klassifikation krankheitsrelevanter Sätze, da sich die annotierten Publikationen ausschließlich mit Transkriptionsfaktoren in roten Blutkörperchen beschäftigen.

Statt Worten oder Stammformen konnten die semantischen Felder der Sätze als Attribute verwendet werden. Mit nur etwa 130 in den Kurzzusammenfassungen vorkommenden Feldern eignen sie sich deutlich besser

für ein Training des SVM-Modells als die knapp 50 000 verschiedenen Konzepte, von denen nur ein Bruchteil innerhalb des annotierten Trainingsdatensatzes vorkommt.

Durch die Verwendung des SVM-Filters konnte eine Präzision von 92 % bei einer Vollständigkeit von annähernd 50 % erreicht werden. Ohne Verwendung des SVM-Filters wurden bei gleicher Präzision nur zwischen 30 und 40 % Vollständigkeit erzielt (siehe Tabellen 3.9 und 3.12 auf Seite 61). In der Kombination mit den anderen drei Filtern – der Mindestbewertungszahl, der Mindestzahl an Kollokationen und der Entfernung von Negationen – konnte die Präzision bis auf 97 % gesteigert werden, was aber aufgrund der sehr niedrigen Vollständigkeit von 20 % nur eine theoretische Rolle spielt.

Die durch den SVM-Filter entfernten Sätze entsprechen häufig denen, die auch durch höhere Mindestzahl von Kollokationen entfernt werden. In beiden Fällen werden dabei aufgrund von Labormethoden oder anderen seltenen Kollokationen hergestellte falsche Zuordnungen entfernt. Bei der Verwendung des SVM-Filters geschieht dies jedoch über den semantischen Kontext, also unabhängig von der Anzahl der vorgefundenen Kollokationen. Das hat zur Folge, dass die Präzision bei der Verwendung des SVM-Filters ähnlich hoch ist wie bei einer Mindestanzahl von drei oder vier Kollokationen, dafür aber eine deutlich bessere Vollständigkeit erreicht wird. Die Klassifikation von Sätzen aufgrund ihrer semantischen Felder erlaubt es in vielen Fällen, zwischen selten in der Literatur vertretenen, aber richtigen Zusammenhängen, und seltenen zufälligen Kollokationen zu unterscheiden.

Limitationen und Alternativen

Trotz einem niedrigen Verhältnis von nur 1,7 Zuordnungen je Satz bleibt das Problem der hohen Konnektivität bei der Verwendung von Konzepten erhalten, wenn auch in verminderter Form (siehe Abbildung 4.2). Mit zehn automatisch zugeordneten Krankheitskonzepten je Enzymklasse liegt diese um den Faktor vier unter der Konnektivität der auf MESH-Begriffen basierenden Zuordnungen und erleichtert eine Analyse des Netzwerks aus Enzymen und Krankheiten. Ein Grund für die hohe Anzahl von zugeordneten Krankheitskonzepten – bei restriktiven Filtern immer noch mehr als 1 400 – liegt in der hohen Zahl sehr ähnlicher Konzepte. Eine

Zusammenfassung benachbarter Konzepte, wie sie bei der Verwendung der MESH-Hierarchie erfolgte, ist aufgrund der beschriebenen Probleme des semantischen Netzwerks der UMLS-Ontologie schwierig.

Eine Abbildung der Konzepte auf das mit den wenigsten Schritten im semantischen Netzwerk zu erreichende MESH-Konzept ist möglich, aber mit einer Präzision von 65 % bei einer Vollständigkeit von 65 % zu fehlerträchtig [19]. Eine alternative Lösung ist die Gruppierung von Konzepten aufgrund ihrer ähnlichen Bezeichnungen: dem kürzesten gemeinsamen Konzeptnamen. So würden die Konzepte *migraine* und *common migraine* als *migraine* zusammengefasst [116]. Da ein Vorteil von Konzepten in der Unabhängigkeit von Schreibweisen und Formulierungen besteht, erscheint dies eher kontraproduktiv. Auf eine Zusammenfassung der Konzepte wurde daher zugunsten einer höheren Präzision verzichtet.

Ein weiterer Grund für die hohe Konnektivität sind falsch positive Zuordnungen, die trotz aller Filter entstehen können. Innerhalb der in dieser Arbeit manuell annotierten Sätze haben klinische Beschreibungen von Blutwerten, Enzymaktivitäten und klinischen Versuchen einen hohen Anteil an den hergestellten Zuordnungen (siehe Tabelle 4.3). Inwiefern diese Symptom oder Ursache einer Krankheit darstellen, ist nicht immer anhand eines einzelnen Satzes auszumachen. Auch Beschreibungen verschiedener Bakterienstämme über Unterschiede einer Enzymsequenz können zu falsch positiven Zuordnungen führen:

> — *Patterns of ribosomal dna polymorphism were examined to compare* carboxylesterase b *type b1 strains and b2 strains of escherichia coli isolated from* extra-intestinal infections.

Auf solchen Sätzen basierende Zuordnungen unterscheiden sich auch in ihren semantischen Feldern nicht von eindeutigen Zuordnungen und stellen damit trotz aller Filter eine mögliche Fehlerquelle dar.

Bewertung der Zuordnungen von Enzymklassen zu Krankheitskonzepten

Bei der entwickelten satzbasierten Extraktion von Krankheitskonzepten und ihrer Zuordnung zu Enzymklassen handelt es sich um eine präzise Methode, Datenbanken mit neuen Informationen zu erweitern und

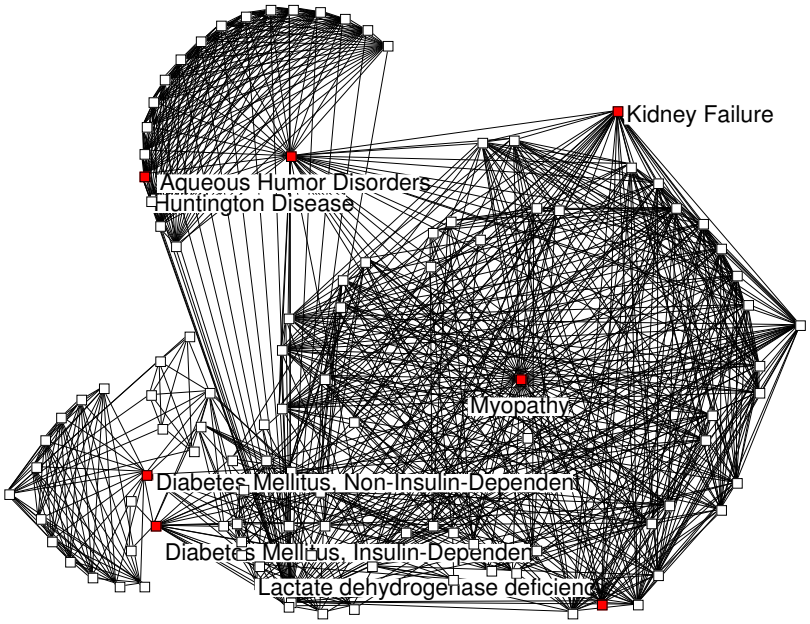


Abbildung 4.2 · Darstellung der Konnektivität von Krankheiten, denen Enzyme aus der Glykolyse oder Gluconeogenese zugeordnet wurden. Eine Kante zwischen zwei Krankheitskonzepten besteht genau dann, wenn beide Knoten mindestens eine Enzymklasse gemeinsam haben. Zur Verbesserung der Übersicht sind nur einige Knoten mit Namen versehen.

Tabelle 4.3 · Konzepte und semantische Felder in den Sätzen des Trainingskorpus mit einem Zusammenhang zwischen Krankheitskonzept und Enzymklasse.

Rang	Konzepte	Semantische Felder
1	patient	Disease or Syndrome
2	activity	Amino Acid, Peptide, or Protein
3	deficiency	Enzyme
4	control	Functional Concept
5	enzymes	Qualitative Concept
6	normal	Finding
7	disease	Laboratory Procedure
8	levels	Pharmacologic Substance
9	cases	Patient or Disabled Group
10	study	Quantitative Concept

Lücken in der manuellen Annotation schnell zu schließen. Die in dieser Arbeit entwickelte Methode lässt sich problemlos auf andere Textkorpora übertragen; ebenso ermöglicht sie die Herstellung anderer Zuordnungen als Krankheiten und Syndrome, solange diese als semantisches Feld in der UMLS-Ontologie vertreten sind. So ist für die Extraktion von pharmazeutischen Wirkstoffen oder Symptomen wie Nebenwirkungen von Medikamenten keine Änderung am Programm oder der Datenbank notwendig, nur das Training des SVM-Filters erfordert die erneute Annotation eines kleinen Textkorpus.

4.3 Hypothesen & Netzwerkanalyse

Auch wenn die manuelle Analyse der Zuordnungen mittels der implementierten Visualisierungen (siehe Abschnitt 3.5 auf Seite 70) die zuverlässigste Methode darstellt, eventuelle neue Erkenntnisse aus den gewonnenen Daten zu erhalten und unerwartete Zusammenhänge aufzudecken, kann dies aufgrund der Menge an Informationen zu einem schwierigen Unterfangen werden. Hilfreich sind bei der Auswertung daher Verfahren, welche die Daten eingrenzen beziehungsweise die Aufmerksamkeit auf einzelne Teilbereich lenken.

4.3.1 Verteilung von Krankheiten und Enzymklassen

Allgemein auffällig ist der starke absolute Anteil der Hydrolasen unter denen den Krankheiten zugeordneten Enzymklassen (siehe Tabelle 3.13 und Abbildung 3.6 auf Seite 63). Der Anteil der Hydrolasen am humanen Proteom liegt mit 10,9 % über dem der Oxidoreduktasen (8,6 %), was die Unterschiede zwischen beiden Klassen erklären mag – allerdings liegt der Anteil der Transferasen mit 13,1 % noch höher [79]. Inwieweit diese Häufung an der Ähnlichkeit der Reaktion der Hydrolasen und Transferasen[3], der oftmals geringen Substratspezifität [77] oder an einem anderen Grund wie einer Überrepräsentation in wichtigen metabolischen Pfaden liegt, lässt sich mit den vorliegenden Daten nicht feststellen.

[3] Im Prinzip können hydrolytische Enzyme als Transferasen klassifiziert werden, wobei eine Gruppe auf ein Wassermolekül übertragen wird.

Die Liste der Enzymklassen mit der höchsten Anzahl an zugeordneten Krankheiten (siehe Tabelle 3.14 auf Seite 64) führen mit je mehr als hundert Krankheitskonzepten Renin und Peptidyl-Dipeptidase A (oder ACE für *Angiotensin converting enzyme*) an. Das von beiden Enzymen gebildete Angiotensin II spielt bei der Regulation des Blutdrucks und der Flüssigkeitsbalance eine wichtige Rolle. Veränderungen des Blutdrucks, auch aufgrund von medikamentösen Nebenwirkungen, werden für einer Vielzahl von Krankheiten beschrieben, was die hohe Konnektivität der beiden Enzymklassen erklärt.

Die pankreatische Elastase wird mit 97 Krankheitskonzepten in Verbindung gebracht, darunter die Pankreatitis, zystischer Fibrose und Diabetes Mellitus; ebenso finden sich verschiedene andere Elastasen aufgrund ihrer Beteiligung an entzündlichen Erkrankungen und Immunreaktionen unter den Enzymklassen mit der höchsten Anzahl zugeordneter Krankheiten. Ebenfalls eine Rolle bei chronischen Entzündungen sowie Zivilisationskrankheiten wie Fettleibigkeit und Herzinfarkten spielt die NO-Synthase.

Ein ähnliches Bild findet sich bei den Krankheitskonzepten mit der höchsten Zahl zugeordneter Enzymklassen: Formen von Diabetes finden sich in den oberen Rängen, genau wie das damit assoziierte Nierenversagen. Der zweite große Block an Krankheiten wird von Formen der Arthritis gebildet, einer ebenso multifaktoriellen Erkrankung mit einer Vielzahl von Ausprägungen, Risikofaktoren und Symptomen. Für die aktuelle Forschung relevanter als die weitgehend bekannten Fakten über die beschriebenen Volkskrankheiten und hochkonnektierten Enzyme sind unerwartete spezifischere Zuordnungen, die mit den im folgenden diskutierten Verfahren aufgedeckt werden konnten.

Krankheiten und metabolische Pfade

Die Untersuchung der Annotation von jeweils drei Enzymklassen auf fehlende Krankheitszuordnungen (siehe Abschnitt 2.4.3 auf Seite 37 und Abbildung 4.3 für ein Beispiel) resultierte in mehreren Schwierigkeiten. Unter den verwendeten stringenten Parametern der Zuordnung fanden sich wenige Kombinationen von Enzymklassen einer Substrat-Produkt-Kette, bei der dem zweiten Enzym der Kette grundlegend andere Krankheitsbegriffe zugeordnet wurden. In der überwiegenden Mehrzahl handelte es sich in um Varianten des gleichen Krankheitsbegriffs wie eine

wechselnde Annotation mit Diabetes Mellitus und dem übergeordneten Begriff Diabetes, eine Folge der bereits diskutierten Zuordnung mit ähnlichen Krankheitskonzepten. In anderen Fällen resultieren Liganden wie ATP oder NADH in zu allgemeinen Substrat-Produkt-Ketten. Hier wäre eine Unterscheidung zwischen Substrat und Kosubstrat oder alternativ eine Filterung niedermolekularer Substrate notwendig, um weniger stringente Parameter bei der Zuordnung anwenden zu können.

Die Einbeziehung von Enzymklassen, die der KEGG-Datenbank zufolge nicht im menschlichen Organismus vorkommen, bringt zwar keine weiteren funktionellen Zusammenhänge, kann aber bei der Hypothesengenerierung für diagnostische und therapeutische Ansätze hilfreich sein. So wird für das Reye-Syndrom, einer meist durch die Aspirinbehandlung eines grippalen Effekts hervorgerufene Erkrankung von Kindern, das pflanzliche Enzym Putrescin N-Methyltransferase EC 2.1.1.53 gefunden. Es verbindet die mit der Krankheit annotierten Enzymklassen Ornithincarboxylase (EC 4.1.1.17) und die Amin-Oxidase (EC 1.4.3.6) [57]. Die Folgen der durch die Aspiringabe inhibierten β-Oxidation von langkettigen Fettsäuren – mitochondrialer Stress und Zelltod – werden durch eine höhere Polyaminsynthese, an der die Ornithincarboxylase beteiligt ist, noch verstärkt [103, 171]. In Pflanzen gibt es dagegen mit der Putrescin N-Methyltransferase einen alternativen Stoffwechselweg zur Umwandlung von Putrescin, dem Produkt der Ornithincarboxylase.

Ein gutes Beispiel für die Probleme der Namensgebung von Enzymklassen stellt eine andere gefundene Kombination von drei Enzymklassen dar, die mit der Galaktosämie (OMIM-Eintrag 230400) annotiert wurde: Bei der klassischen Galaktosämie liegt eine Störung der UDP-Glukose-Hexose-1-Phosphat Uridylyltransferase (Synonym GALT, EC 2.7.7.12) vor, einem Enzym des Leloir-Stoffwechselweges (siehe Abbildung 4.4). Die re-

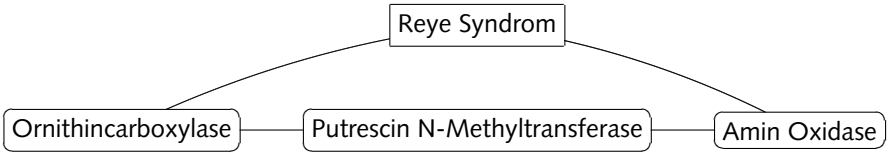


Abbildung 4.3 · Beispiel für die in metabolischen Pfaden gesuchten Dreiergruppen von Enzymen. Nur die Putrescin N-Methyltransferase ist nicht dem Reye-Syndrom zugeordnet.

sultierende Akkumulation von Galaktose-1-Phosphat führt unter anderem zu Nieren- und Leberversagen [118]. Ebenfalls an diesem Stoffwechsel beteiligt – und die zweite Enzymklasse der Substrat-Produkt-Kette – ist die UDP-Galaktose-4-Epimerase (EC 5.1.3.2). Ein Zusammenhang zwischen der Epimerase und der Galaktosämie wird allerdings nur einmal innerhalb des verwendeten Textkorpus beschrieben. Daher wurde der Zusammenhang nicht gefunden und führte zu einer Lücke in der Annotation.

Interessant ist diese Substrat-Produkt-Kette dennoch aufgrund der dritten Enzymklasse, da sie ein gutes Beispiel für die Probleme der Namensgebung von Enzymen darstellt. Die UTP-Hexose-1-Phosphat Uridylyltransferase (EC-Nummer 2.7.7.10) wurde in der in dieser Arbeit verwendeten Version der BRENDA-Datenbank noch fälschlicherweise unter dem Synonym *Galaktose 1-Phosphat Uridylyltransferase* geführt, ein andere Name für die UDP-Glukose-Hexose-1-Phosphat Uridylyltransferase (GALT, EC-Nummer 2.7.7.12) – dem ersten an der Galaktosämie beteiligten Enzym[4] (siehe Tabelle 4.4 für eine Namensübersicht). Auch im verwendeten Textkorpus werden die EC-Nummern 2.7.7.10 und 2.7.7.12 austauschbar verwendet, dazu oft noch mit den empfohlenen Namen beider Enzymklassen gleichzeitig [62]. In der KEGG-Datenbank wird die UTP-Hexose-1-Phosphat Uridylyltransferase nicht als humanes Enzym geführt, und in der OMIM-Datenbank findet sich kein Hinweis auf einen möglichen Zusammenhang zwischen Enzym und Galaktosämie. Dabei gibt es Publikationen, dass das Enzym in Galaktosämie-Patienten Galaktose-1-Phosphat und UTP zu UDP-Galaktose und PP_i umwandelt und dadurch die Transfereaktion der GALT umgeht. Diese Reaktion ist aufgrund ihrer niedrigen spezifischen Aktivität in Kontrollen von nicht erkrankten Personen nicht zu finden [33]. Fehler wie diese sind ohne eine manuelle Durchsicht der Literatur praktisch nicht zu entdecken. Allerdings ist es durch die Verfahren des *Text Mining* möglich, solche Widersprüche und Lücken in der Annotation einfach aufzudecken und so die manuelle Auswertung zu unterstützen.

Die Anwendung weniger stringenter Parameter bei der Krankheitszuordnung führten aufgrund der höheren Zahl falscher Zuordnungen zu einem drastischen Anstieg der zu untersuchenden Substrat-Produkt-Ketten,

[4] In der aktuellen Version (April 2004) wurde der Fehler korrigiert.

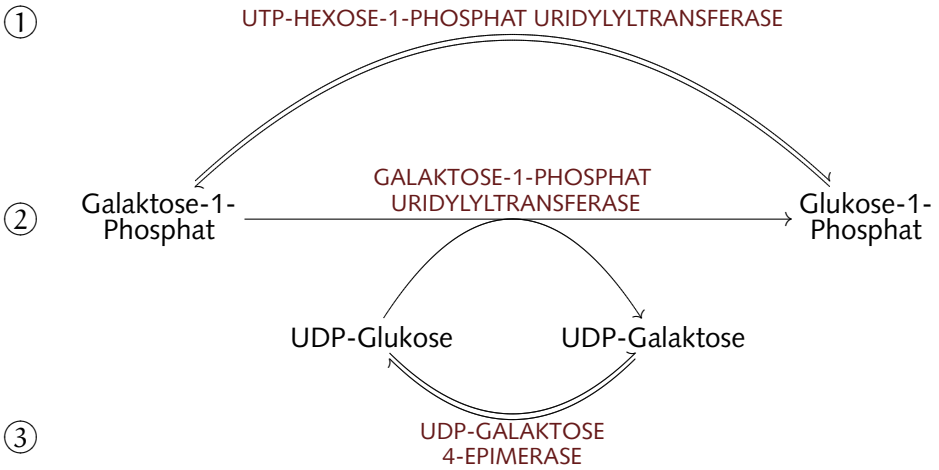


Abbildung 4.4 · Auszug aus dem Leloir-Stoffwechselweg zur Umwandlung von β-D-Galaktose zu Glukose-1-Phosphat. Schritte 2 und 3 zeigen die Umwandlung von Galaktose-1-Phosphat zu Glukose-1-Phosphat unter Verwendung von UDP-Glukose; Schritt 1 den in Zellen von Galaktosämie-Patienten gefundenen alternativen Pfad.

Tabelle 4.4 · Empfohlene Enzymnamen und ihre Synonyme in der BRENDA-Datenbank für zwei an der Galaktosämie beteiligte Enzymklassen.

EC-Nummer	Empfohlener Name	Synonyme
2.7.7.12	UDP-glucose-hexose-1-phosphate uridylyltransferase	Galactose-1-phosphate uridylyltransferase, UDPGlucose:α-D-galactose-1-phosphate uridylyltransferase, GALT
2.7.7.10	UTP-hexose-1-phosphate uridylyltransferase	UTP:α-D-galactose-1-phosphate uridylyltransferase, UDP-Galactose pyrophosporylase

was eine sinnvolle Auswertung der Krankheitszuordnungen verhinderte. Für eine Reduktion der Anzahl an gefundenen Substrat-Produkt-Ketten sind zusätzliche Informationen notwendig. Daten wie die subzelluläre Lokalisation und Hinweisen auf die Gewebe, in denen die Enzymklassen im menschlichen Organismus vorkommen, helfen dabei ebenso wie Hinweise zur Richtung und Reversibilität der katalysierten Reaktion unter physiologischen Bedingungen. Keine dieser Daten ist in der KEGG-Datenbank vorhanden, könnten jedoch der aktuellen Version der BRENDA-Datenbank entnommen und dazu genutzt werden, eine Erweiterung der Analyse über die artifiziellen Grenzen der metabolischen KEGG-Karten hinaus zu realisieren. Wie die Überschneidung der metabolischen Pfade in Tabelle 3.17 auf Seite 67 verdeutlicht, handelt es sich nicht um geschlossene Systeme; in Fällen wie der Glykolyse und Gluconeogenese sind annähernd 70 % der beteiligten Enzymklassen an weiteren metabolischen Pfaden beteiligt.

Sequenzähnlichkeiten krankheitsrelevanter Enzymklassen

Eine Möglichkeit, Enzymklassen und ihre Krankheitszuordnungen über ihren metabolischen Zusammenhang hinaus miteinander zu vergleichen, ist die Analyse ihrer Sequenzähnlichkeit. Enzyme unterschiedlicher EC-Nummern mit einer hohen Sequenzähnlichkeit lassen Spekulationen über ähnliche Reaktionsmechanismen [11], Interaktionspartner [31] oder Proteinstrukturen [61] zu. Bei der Zuordnung eines gemeinsamen Krankheitskonzepts zu Enzymklassen mit einer hohen Sequenzähnlichkeit stellt sich die Frage, inwieweit die Sequenzähnlichkeit in der Literatur beschrieben ist und ob die damit zusammenhängende Enzymfunktion kausal an der Krankheit beteiligt ist. Die in Tabelle 3.18 auf Seite 69 gezeigten Zusammenhänge liefern dazu einige Beispiele:

- Die Cardiomyopathie ist eines der Symptome der Glykogenspeicherkrankheit Typ II (Pompe-Krankheit, OMIM-Eintrag 232300), die auf einem Defekt oder dem Fehlen des α-Glucosidase-Gens beruht. Ein Defekt der 4-α-Glucanotransferase führt zu der Glykogenspeicherkrankheit Typ III und resultiert unter anderem ebenfalls in einer Cardiomyopathie (Forbes-Krankheit, OMIM-Eintrag 232400). Beide Enzymklassen sind, wie auch die Pullulanase, Teil der Glykosid-

Hydrolase-Familie 13 in der *CAZy*-Datenbank, welche kohlenhydratbindende und -modifizierende Enzyme aufgrund von Ähnlichkeiten in ihrer Struktur klassifiziert [40]. Pullulanase selber ist aber nicht Bestandteil des menschlichen Genoms, die Zuordnung zur Cardiomyopathie erfolgte aufgrund des Synonyms *debranching enzyme*, welches es mit der 4-α-Glucanotransferase gemein hat. Ein Zusammenhang mit der Krankheit ist hier also nicht gegeben.

- Die periphere Neuropathie führt zu Taubheitsgefühlen und Schmerzen in den Extremitäten und hat eine Vielzahl von Ursachen. Neben Formen von Diabetes, Metallvergiftungen und den Folgen von Alkoholismus wird sie vor allem durch Ernährungsmängel verursacht, die zu einer Thiamindefizienz (Vitamin B1) führen. Sowohl Pyruvatdehydrogenase als auch Transketolase gehören zu den thiaminabhängigen Enzymklassen. Eine Defizienz der Pyruvatdehydrogenase kann unter anderem zu einer peripheren Neuropathie führen [32]. Eine Änderung der Thiaminbindungseigenschaften der Transketolase bei dem Wernicke-Korsakoff-Syndrom führt insbesondere bei Alkoholikern mit einem Thiaminmangel ebenfalls zu einer peripheren Neuropathie (OMIM-Eintrag 277730). Der Zusammenhang zwischen beiden Enzymklassen konnte aufgrund der Sequenzähnlichkeit ohne explizite gemeinsame Nennung innerhalb des Textkorpus identifiziert werden.

- Etwas komplizierter ist die Beschreibung der ‚Xanthin- und Aldehydoxidase-Defizienz', auch wenn der Name der Krankheit bereits andeutet, dass die Beteiligung beider Enzyme keine Überraschung ist – sie verdeutlicht aber die prinzipiell mit dieser Methode zu erreichenden Ergebnisse. Bei der Xanthin-Dehydrogenase/Oxidase handelt es sich um ein multifunktionelles Enzym, welches im menschlichen Metabolismus die letzten zwei Schritte des Purinkatabolismus katalysiert [126]. Ein Defekt in diesem Enzym führt zur Xanthinurie Typ I (OMIM-Eintrag 278300), einer Krankheit bei der große Mengen des schwer löslichen Xanthins entweder mit dem Urin ausgeschieden werden oder sich ablagern. Bei der Xanthinurie Typ II (OMIM-Eintrag 603592 und [119]) liegt zusätzlich noch eine Defizienz der Aldehyddehydrogenase vor, wodurch Allopurinol – ein In-

hibitor der Xanthin-Oxidase – nicht mehr zu Oxypurinol oxidiert werden kann.

Die Aldehyddehydrogenase verfügt über ein der Xanthin-Oxidase nahezu identisches Elektronentransportsystem, hat aber eine höhere Affinität zu Aldehyden statt Purinen; beide Enzyme werden von den gleichen Antikörpern erkannt. Mittlerweile wird vermutet, dass es sich bei der Typ II-Variante nicht um strukturelle Defekte der beiden Enzyme handelt, sondern der Mechanismus zur Insertion eines Kofaktors die Ursache ist [140].

Das letzte Beispiel verdeutlicht, dass eine Untersuchung der Enzymcluster mit einer hohen Sequenzähnlichkeit in Bezug auf die den Enzymklassen zugeordneten Krankheiten nach Abschluss der Arbeit von Christian aus dem Spring sinnvoll ist.

4.3.2 Topologie des Enzym- und Krankheitsgraphen

Eine andere Betrachtungsweise entsteht durch die Analyse der Netzwerkstruktur, sowohl für den Graphen aus Krankheitskonzepten als auch für die Enzymklassen (siehe Abschnitt 2.4.3 auf Seite 38). Die Ermittlung von Werten wie dem mittleren Abstand von Knoten, der Gruppierungseigenschaften oder der Verteilungsfunktion der Knotengrade ermöglicht eine genauere Charakterisierung der bisher beschriebenen Grapheneigenschaften. Dadurch ist ein Vergleich mit anderen biologischen und sozialen Netzwerken möglich, was Rückschlüsse auf die Art der Zuordnungen erlaubt [59]; nach Strogatz beeinflusst die Netzwerkstruktur immer die Funktion.

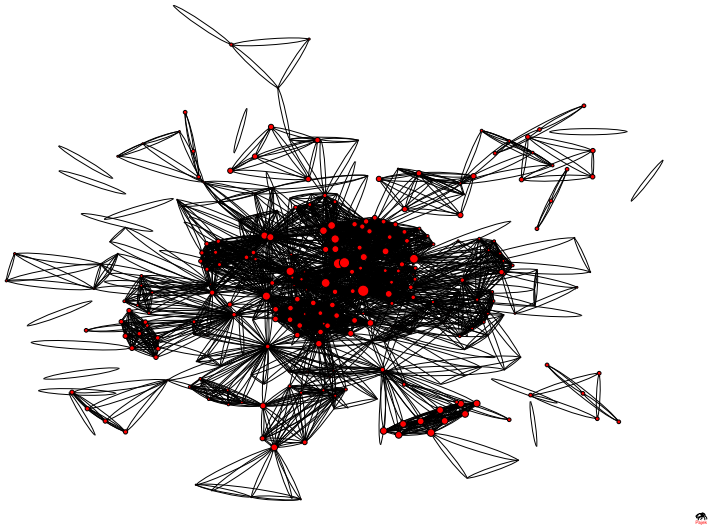
Die vorgefundenen Eigenschaften charakterisieren die Topologie beider Netzwerke als *Small World*-Graphen, wie sie auch für metabolische Graphen [59], soziale Netzwerke [108] oder die Neuronenstruktur von *C. elegans* bekannt ist [158]. Die Gruppierung ähnlicher Krankheitskonzepte aufgrund der ihnen gemeinsamen Enzymklasse führt zu einem höheren Gruppierungskoeffizient, als dies bei zufälligen Verbindungen zu erwarten wäre. Gleichzeitig lassen allgemeine Krankheitskonzepte, die mit einer hohen Zahl von Enzymklassen verbunden sind, die für einen *Small World*-Graphen notwendigen Abkürzungen durch das Netzwerk entstehen. Dies

gilt entsprechend für das Netzwerk aus Enzymklassen, hier sind es hochkonnektive Enzyme wie Proteinkinasen, die zu einem geringen mittleren Abstand der Knoten führen.

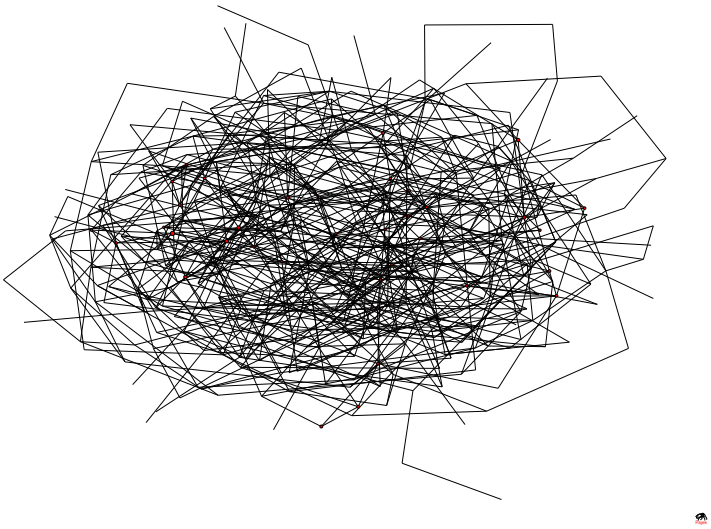
Diese zentralen Knoten sind es, die zu der größenunabhängigen Verteilung der Knotengrade führen. Die Architektur der untersuchten Graphen unterscheidet sich deutlich von der eines Graphen mit zufälliger Kantenverteilung (siehe Abbildung 4.5) und entspricht den sogenannten *Scale Free*-Graphen. Diese basieren auf dem Modell von wachsenden Netzwerken, bei denen neu hinzukommende Knoten bevorzugt Kanten zu bereits hochkonnektierten Knoten ausbilden. Unabhängig von der Netzwerkgröße finden sich alle Knotengrade wieder. Aufgrund der bevorzugten Anbindung werden die Graphen von *Hubs* dominiert, zentrale Knoten mit einer hoher Konnektivität [14].

In semantischen Netzwerken ist die Bildung von zentralen Knoten bekannt [146]. Diese entstehen unabhängig von ihrer syntaktischen Klasse vermutlich aufgrund der Art und Weise, mit der ein Vokabular gebildet wird. Alte Begriffe bilden die Basis eines Wörterbuchs. Werden im Laufe der Zeit neue Konzepte hinzugefügt, so erweitern sie komplexe oder zentrale, ältere Begriffe. Als zentrale Knoten entstehen dabei verbindende Konzepte, ganz analog zu einem zentralen Metaboliten wie ATP im Stoffwechsel [59, 82]. Diese Eigenschaft zeigt sich auch für die Krankheitsbegriffe der UMLS-Ontologie (siehe Abbildung 4.6), bei der die Konzepte nicht nach ihrem Namen, sondern ihrer laufenden Identifikationsnummer sortiert wurden. Ältere Konzepte mit einer niedrigeren Nummer sind im Schnitt mit mehr Krankheiten verbunden als solche, die der Ontologie erst kürzlich hinzugefügt wurden. Dies mag daran liegen, dass bereits länger bekannte Krankheiten genauer untersucht worden sind als die erst in den letzten Jahren als krankheitsrelevant bekannt gewordenen Konzepte. Ebenso werden einer allgemeinen Ontologie als erstes die wichtigsten und bekanntesten Konzepte hinzugefügt, bei denen es sich im Krankheitsfall um gut untersuchte Volkskrankheiten mit vielen Zitaten in der Literatur handeln dürfte. Inwieweit die festgestellten Eigenschaften des Graphen daher tatsächlich Eigenschaften der Krankheiten sind ist schwer zu bestimmen, da dazu eine Ontologie benötigt würde, die keine *Scale-Free*-Architektur aufweist.

Für den Enzymgraph lassen sich ähnliche Überlegungen anstellen. Zentrale Knoten in metabolischen Netzwerken lassen sich durch die Evoluti-



(a) Krankheitsgraph für 10 % des Textkorpus



(b) Zufälliger Graph mit der gleichen Anzahl an Knoten und Kanten

Abbildung 4.5 · Darstellung des Krankheitsgraphen (a) und eines zufälligen Graphen (b) mit der gleichen Anzahl an Kanten und mittlerer Konnektivität mit *Pajek*. Um die Visualisierungen anschaulich zu halten war eine Reduktion des Textkorpus auf 10 % notwendig. Die Größe der Knoten entspricht ihrem Gruppierungskoeffizienten.

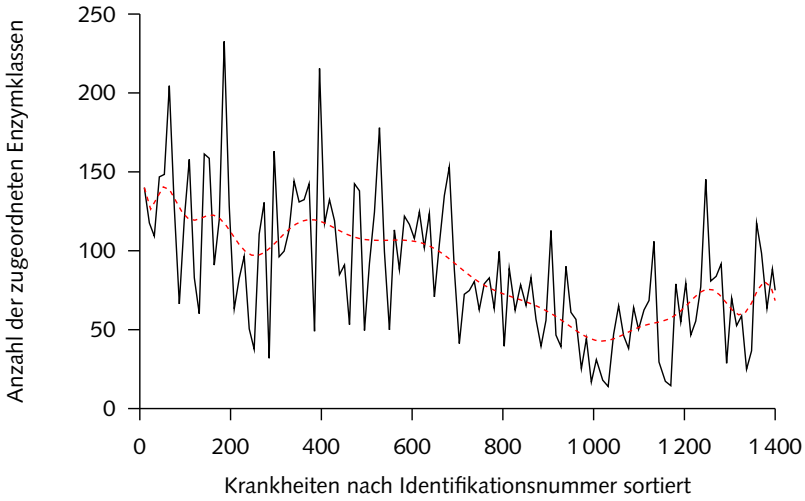


Abbildung 4.6 · Aufgetragen ist die mittlere Zahl von Enzymklassen, die je zehn Krankheitskonzepten der UMLS-Ontologie zugeordnet werden konnten. Die Krankheitskonzepte sind nach ihrer Identifikationsnummer aufsteigend sortiert, erst kürzlich der Ontologie hinzugefügte Konzepte finden sich auf der rechten Seite des Graphen.

on erklären – solange kein hochkonnektiertes Schlüsselenzym von einer Mutation betroffen ist, sind solche Netzwerke äußerst robust gegenüber zufälligen Mutationen [81, 94]. Bei Netzwerken von Proteindomänen entstehen sie vermutlich aufgrund der Expansion bestehender Domänen, um so insbesondere in höheren Organismen eine stärkere Komplexität der Regulation und Transduktion zu ermöglichen [167]. Ein ähnlicher Hintergrund ist auch für den vorliegenden Enzymgraphen denkbar: evolutionär ältere Enzyme sind zentrale Bestandteile des Stoffwechsels, verbinden unterschiedliche Stoffwechselwege miteinander [59] und könnten dadurch an mehr Krankheiten beteiligt sein. Aufgrund der Gruppierung zu Enzymklassen mit einer sehr unterschiedlichen Anzahl von Enzymsequenzen fällt es jedoch schwierig, diese Vermutung zu überprüfen. Notwendig hierfür wäre eine Zuordnung der Krankheiten zu Enzymsequenzen sowie deren Mutationen. Diese Information findet sich aber nur in den seltensten Fällen in den Publikationen wieder.

Sowohl die Eigenschaften des Krankheits- als auch des Enzymgraphen werden durch nicht-biologische Einflüsse verzerrt. Eine Unterscheidung,

was biologischer Hintergrund und was Einfluss der Ontologie oder der Einteilung in Enzymklassen ist, bleibt schwierig. Dies gilt auch für die publizierten Eigenschaften anderer biologischer Netzwerke, die durch ähnliche Umstände beeinträchtigt werden. Unabhängig von der biologischen Signifikanz muss die beobachtete Topologie bei der Entwicklung einer Bewertungsfunktion für die Zuordnung von Enzymklassen zu Krankheiten berücksichtigt werden, wenn diese auf der Anzahl der Verknüpfungen oder Nennungen in der Literatur basieren soll.

4.3.3 Auswertung der Subgraphen

Bei der in Abschnitt 2.4.3 auf Seite 39 beschriebenen Suche nach nicht vollständig verbundenen Subgraphen ging es um die Aufdeckung impliziter Gemeinsamkeiten zwischen Krankheitskonzepten. Ähnlich zum beschriebenen Swanson-Modell sollten Zusammenhänge zwischen Krankheiten untersucht werden, die keiner gemeinsame Enzymklassen zugeordnet waren. Ausgewertet wurden dazu Subgraphen, in denen die beiden Konzepte implizit in Verbindung gebracht wurden (siehe Abbildung 4.7). Bei der Suche nach solchen implizit miteinander verknüpften Konzepten gab es zwei Eigenschaften, die gegeneinander abgewogen werden mussten: Die Anzahl an verbindenden Konzepten sowie die Anzahl der Kanten des untersuchten Subgraphen.

Werden die nicht verbundenen Konzepte (Krankheiten A und C in Abbildung 4.7) durch mehrere miteinander verbundene Konzepte (Krankheiten B und D) in Verbindung gebracht, so steigt die Konfidenz in einen möglichen Zusammenhang. Dies gilt allerdings nicht unbegrenzt: Bei einer sehr hohen Zahl von indirekten Verbindungen erscheint es unwahr-

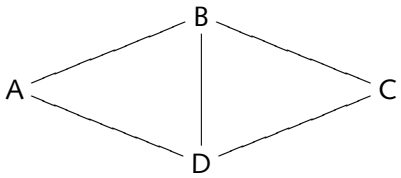


Abbildung 4.7 · Beispiel für einen gesuchten Subgraphen aus vier Krankheitskonzepten, bei dem nur die Krankheiten A und C nicht miteinander verbunden sind.

scheinlich, dass ein Zusammenhang besteht und noch nicht entdeckt wurde. Ebenfalls berücksichtigt werden musste die Konnektivität des Subgraphen. Allgemeine Krankheitskonzepte, die mit fast allen anderen Krankheiten eine Enzymklasse gemeinsam haben, sollten ein geringeres Gewicht haben.

Daher wurden bevorzugt Hypothesen untersucht, bei denen die beiden Krankheiten Bestandteil von zwei bis fünf Subgraphen mit möglichst geringer durchschnittlicher Konnektivität waren. Abbildung 3.9 auf Seite 72 zeigt dafür ein Beispiel. Es bringt die Phenylketonurie mit einer Form der Dystonie in Verbindung. Bei der klassischen Phenylketonurie liegt ein Defekt der Phenylalanin-Hydroxylase (Phenylalanin 4-Monooxygenase, EC 1.4.16.1) vor. Dadurch kann Phenylalanin nicht in Tyrosin umgewandelt werden und reichert sich mit seinen Zwischenprodukten ohne eine entsprechende Diät im Blut an, was unter anderem eine schwere geistige Retardierung und Muskelkrämpfe zur Folge hat [49]. Damit verwandt ist die Hyperphenylalaninaemie oder auch atypische Phenylketonurie. Hierbei kann aufgrund eines Defekts der GTP-Cyclohydrolase I (EC 3.5.4.16) oder der Dihydropteridin-Reduktase kein Tetrahydrobiopterin (BH_4) regeneriert oder synthetisiert werden – ein Kofaktor der Phenylalanin 4-Monooxygenase [110]. Bei der Synthese von Neurotransmittern wie Serotonin oder Dopamin durch die Tyrosin-Hydroxylase spielen sowohl BH_4 als auch Tyrosin eine Rolle[5], so dass bei dieser Form der Phenylketonurie neurologische Symptome aufgrund eines Neurotransmittermangels hinzukommen [23]. Dieser Mangel wurde auch in Folge der dietären Behandlung von Phenylketonurie-Patienten beobachtet [27].

Dystonien zeichnen sich durch ungewollte, starke Muskelkrämpfe aus. Die familiäre Form wird autosomal dominant durch eine Mutation des ATP-bindenden Proteins Torsin A vererbt (DYT1, OMIM-Einträge 128100 und 224500). Ein Subtyp beruht dagegen auf einem Defekt der GTP-Cyclohydrolase I (DYT5, OMIM-Eintrag 128230) [104] – mit den gleichen Folgen eines Dopaminmangels wie bei den Phenylketonurien.

Da die UMLS-Ontologie zwischen beiden Formen der Phenylketonurie unterscheidet ist die fehlende Zuordnung der GTP Cyclohydrolase I zur

[5] GTP-Cyclohydrolase I ist das geschwindigkeitsbestimmende Enzym bei der Umwandlung von GTP zu BH_4, dem Kofaktor der Tyrosin-Hydrolase, welches wiederum das geschwindigkeitsbestimmende Enzym bei der Dopamin-Synthese ist [78].

klassischen Form der Erkrankung richtig. Innerhalb des Textkorpus fand sich der Hinweis auf die Dopamin-abhängige Variante der Dystonien und deren Zusammenhang mit BH_4 und der Phenylketonurie nur innerhalb des Volltextes, jedoch nicht in den ausgewerteten Kurzzusammenfassungen. Hier wäre ein größerer Textkorpus beziehungsweise der Zugriff auf die vollständigen Publikationstexte notwendig.

Unter den Subgraphen mit niedriger Konnektivität fand sich die Galaktosämie unter der Beteiligung der EC-Nummern 2.7.7.10 und 2.7.7.12, welche bereits bei der Auswertung der metabolischen Pfade diskutiert wurde (Abschnitt 4.3.1); ein Hinweis darauf dass die Suche nach nicht vollständig verknüpften Subgraphen zusätzliche Informationen zu anderen Auswertungsmethoden liefern kann.

4.4 Fazit

Erweiterungsmöglichkeiten des entwickelten Programms

Es konnte am Beispiel der Annotation von Enzymklassen mit Krankheitsinformationen gezeigt werden, dass das in dieser Arbeit entwickelte Programm in der Lage ist, biologisch-medizinische Zusammenhänge mit hoher Präzision aus der Literatur zu extrahieren. Die Erweiterung der Methode über die visuelle Repräsentation der Zusammenhänge unterstützt den Einsatz als Forschungswerkzeug zur Entwicklung bisher unbekannter Hypothesen. Der dem Programmpaket zugrunde liegende Ansatz ist dabei nicht zuletzt durch die modulare Implementierung leicht erweiterbar und auf andere Fragestellungen zu übertragen. Die hohe Flexibilität des konzeptbasierten Verfahrens erlaubt es, mit der bestehenden Programmversion neue Fragestellungen zu bearbeiten. So sollten sich mit ähnlicher Präzision und Vollständigkeit auch Informationen über Krankheitssymptome, pharmazeutische Wirkstoffe oder die subzelluläre Lokalisation extrahieren lassen. Geplante Weiterentwicklungen betreffen daher insbesondere solche Erweiterungen des Algorithmus, die diese hohe Flexibilität nicht beeinträchtigen, sowie die Vereinfachung der Benutzerschnittstelle.

Während für die Erweiterung von Datenbanken die Präzision der Zuordnungen höchste Priorität hat, kann in anderen Anwendungen – wie beispielsweise der Suche nach unbekannten Zusammenhängen – der Vollständigkeit eine höhere Bedeutung zukommen. Im Zentrum weiterer Ar-

beiten sollte daher die Entwicklung einer komplexeren Bewertungsfunktion stehen, welche auf den in dieser Arbeit entwickelten Filtern aufbaut und je nach Fragestellung eine Wahl zwischen Präzision und Vollständigkeit erlaubt. Mit einer ihrer geringen Präzision entsprechenden Gewichtung könnten so MESH-Begriffe einbezogen werden, die durch ihren über die Kurzzusammenfassung hinausgehenden Informationsgehalt zur Vollständigkeit beitragen könnten. Ebenso könnte die Einbeziehung von Kollokationen in Nachbarsätzen oder Kurzzusammenfassungen über eine distanzabhängige Bewertungsfunktion die Vollständigkeit weiter erhöhen [165]. Die dabei auftretenden Mehrdeutigkeiten durch den Textbezug über Satzgrenzen hinweg erfordert jedoch aufwändigere linguistische Verfahren [66]. Die semantische Ebene erlaubt dabei mittels Konsistenzprüfungen eine zusätzliche Verbesserung der Präzision. Der Schlüssel hierzu liegt in der Verwendung des semantischen Netzwerks der UMLS-Ontologie, die zur Überprüfung der Konzeptrepräsentation verwendet werden könnte [133, 136].

Zur Verbesserung der Zuordnungen kann schließlich auch eine Syntaxanalyse beitragen. So lassen sich die *part of speech*-Annotation der Konzepte als zusätzliche Attribute für den SVM-Filter verwenden [86] oder in einem *Hidden Markov*-Modell mit der Positionsinformation verknüpfen [43]. Eine weitere Möglichkeit ist die Berücksichtigung spezieller Interaktionsbegriffe, wie sie bei der Erstellung von Proteininteraktionsnetzwerken verwendet werden [100, 111]. Phrasen wie ‚verursacht durch' oder ‚führt zu' deuten fast immer auf einen ursächlichen Zusammenhang hin:

> *— Glycogen storage disease gsd type 1a . . . is caused by a deficiency in glucose-6-phosphatase, the key enzyme in glucose homeostasis. . .*

Um eine Abhängigkeit von existierenden Interaktionsvokabularen zu vermeiden, die zu einer Einschränkung der Anwendungsmöglichkeit führen könnten, sollten Verfahren zur automatischen Erkennung solcher Schlüsselwörter eingesetzt werden [69].

Genauere Aussagen über die extrahierten Zusammenhänge – wie die Stärke einer Aktivierung oder die Richtung einer Inhibierung – sind nur durch deutlich komplexere Verfahren möglich, die über manuell erstellte [113, 117, 164] oder aus Textkorpora [168] gelernte Regelsammlungen und extrahierte Muster auch die Grammatik von Texten aufklären.

Aber auch diese Verfahren erreichen nicht das Abstraktionsvermögen eines menschlichen Lesers. Daher kommt weiterhin der manuellen Annotation eine hohe Bedeutung zu, deren Arbeit das vorliegende Programm bereits jetzt durch die Visualisierung und Parameterwahl unterstützt. Diese Zielsetzung soll in Zukunft durch eine grafische Benutzerschnittstelle weitergeführt werden, welche eine gezielte Abfrage von Enzymklassen, Krankheiten und anderen Suchbegriffen ermöglicht.

Eine Erweiterung der Hypothesengenerierung kann schließlich durch eine engere Integration mit der BRENDA-Datenbank erreicht werden. So kann zum Beispiel bei einer Zuordnung von Symptomen und Nebenwirkungen von Medikamenten zu Enzymklassen ein Vergleich der bekannten Inhibitoren hilfreich sein, Zusammenhänge schneller aufzudecken. Ebenso interessant wäre ein Vergleich von beschriebenen Nebenwirkungen pharmazeutischer Wirkstoffe mit den Substraten und Produkten der Enzymklassen.

Ausblick

Die aufgrund eines Mangels an Volltexten für diese Arbeit notwendige Beschränkung auf Kurzzusammenfassungen resultierte in einer hohen Präzision, die unter Verwendung vollständiger Publikationen nicht ohne weiteres erreichbar gewesen wäre. Das Verhältnis von relevanten Schlüsselwörtern zu irrelevanten Daten ist in Textbereichen wie dem Titel, der Kurzzusammenfassung sowie den Legenden besonders günstig und verbessert die Präzision einer automatischen Auswertung [58]. Eine hohe Informationsdichte alleine garantiert allerdings noch nicht, dass sich alle Informationen aus einem Textbereich wie der Kurzzusammenfassung extrahieren lassen [102]. Insbesondere Informationen zu untersuchten Organismen, Gewebearten oder experimentellen Bedingungen finden sich selten in der Kurzzusammenfassung und machen einen Auswertung der vollständigen Publikation notwendig.

Allerdings verfügt die Biomedizin über kein zentrales Volltext-Archiv, welches die automatische Auswertung dieser Informationen erlauben würde. In vielen Fällen ist der Zugang zu frei verfügbaren Volltexten für computerlinguistischer Methoden ein größeres Hindernis als die Entwicklung eines geeigneten Algorithmus [48]. Seit gut zwei Jahren verstärken sich die Bemühungen, Veröffentlichungen durch Initiativen wie *Open Ac-*

cess[6] im biomedizinischen Umfeld frei verfügbar zu machen. Der Umfang der im *World Wide Web* verfügbaren Publikationen ist in den letzten Jahren sprunghaft angestiegen. Viele Universitätsbibliotheken sind dazu übergegangen, nur noch die elektronischen Ausgaben der Zeitschriften anzuschaffen. Dennoch ist nur ein Bruchteil der Publikationen für eine Auswertung frei zugänglich, dazu meist in proprietären Formaten [48]. Notwendig wären strukturierte Veröffentlichungen, welche nicht nur die unterschiedlichen Abschnitte, Tabellen und Abbildungen einer Publikation klar erkennbar machen, sondern diese auch mit Annotationen versehen. Eine spezielle Annotation für die automatische Textverarbeitung wäre ein erster Schritt, die Verwendung eines geeigneten, kontrollierten Vokabulars aus einer Ontologie eine zusätzliche Hilfe.

Mittelfristig hat die Linguistik das Potential, die zu beobachtende Datenintegration in der Biologie voranzutreiben. Nicht nur die automatische Erweiterung von Datenbanken mit extrahierten Informationen, sondern vor allem die Vereinheitlichung des verwendeten Vokabulars spielen hierbei eine wichtige Rolle. Die zunehmende Anwendung der *Gene Ontology* vereinfacht die konsistente Annotation und dadurch die Integration von Datenbanken wie *Swiss-Prot*, *TrEMBL* und *PIR-PSD* zu einer Informationsquelle wie *Uniprot* [30].

Parallel dazu ist eine Vereinheitlichung der Methoden und Resourcen der biologischen Computerlinguistik notwendig, welche sich ebenfalls abzeichnet. Die Aufnahme von GO und *Snomed* in die UMLS-Ontologie ist hierfür ebenso ein Beispiel wie der Aufbau einer Plattform wie *BioCreative*[7] für den Vergleich linguistischer Methoden in der Biologie. Definierte Textkorpora, Aufgabenstellungen und Bewertungskriterien sollten dabei helfen, ein Bewusstsein für die Eigenheiten biomedizinischer Texte zu schaffen und eine Vergleichbarkeit der unterschiedlichsten Lösungsansätze zu ermöglichen.

[6] http://www.plos.org/about/openaccess.html

[7] Critical Assessment of Information Extraction Systems in Biology: http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

4.5 Zusammenfassung

In dieser Arbeit wurde ein umfangreiches Softwarepaket für verschiedene Aspekte der computergestützten Auswertung wissenschaftlicher Texte entwickelt. Als erste Anwendung wurde die Annotation von Enzymklassen aus der BRENDA-Datenbank mit Krankheitsbezügen durchgeführt. Der hohe Standard dieser von Experten manuell erstellten und gepflegten Enzymdatenbank, die Anforderungen an die Präzision der gefundenen Zuordnungen und die gewünschte Anpassungsmöglichkeit auf unterschiedliche Domänen machte die Entwicklung eines neuen Ansatzes für die Extraktion von Daten aus der wissenschaftlichen Literatur notwendig.

Für dieses Verfahren wurden Methoden aus der statistischen Linguistik mit dem in speziellen Lexika und Ontologien enthaltenen Expertenwissen kombiniert. Basierend auf der UMLS-Ontologie wurde mit Hilfe des von der *Semantic Knowledge Representation Group* zur Verfügung gestellten Programms *MetaMap* eine Konzeptrepräsentation krankheitsrelevanter Phrasen vorgenommen und für die Suche nach Kollokationen mit Enzymbezeichnungen in Einzelsätzen verwendet. Die Identifikation der verschiedenen Enzymbezeichnungen erfolgte hier ebenso wie bei der Auswahl der 105 630 Kurzzusammenfassungen für den Textkorpus durch ein aus der BRENDA-Datenbank erstelltes Enzymlexikon.

Sätze, die Kollokationen aufwiesen, wurden auf vier verschiedene Arten gefiltert: Neben der Mindestzahl an gemeinsamen Nennungen erfolgte eine Filterung basierend auf der internen Bewertungzahl der gefundenen Konzepte durch *MetaMap*. Kollokationen ohne mindestens ein krankheitsrelevantes Konzept mit einer Mindestbewertungszahl von 800 wurden ebenso nicht berücksichtigt wie nur einmalig vorkommende gemeinsame Nennungen. Ein zusätzlicher Filter entfernte alle Sätze mit Negationswörtern, da diese oft zu falschen Zuordnungen führen.

Der letzte Prozessierungsschritt basierte auf den relativen Häufigkeiten von semantischen Feldern, die jedem Konzept in der UMLS-Ontologie zugeordnet sind. Aufgrund der Komplexität der Bewertung dieser semantischen Felder wurde ein Verfahren des computergestützten Lernens auf Basis einer sogenannten *Support Vector Machine* (SVM) entwickelt, die mit 400 positiven und 600 negativen Beispielsätzen trainiert wurde. Dadurch

konnten von der SVM als nicht krankheitsrelevant klassifizierte Sätze vor der Zuordnung entfernt werden.

Insgesamt konnten 524 Enzymklassen der BRENDA-Datenbank 1409 Krankheitskonzepte hinzugefügt werden. Eine manuelle Annotation der Zusammenhänge in 1 500 zufällig ausgewählten Sätzen des Textkorpus aus 105 630 biomedizinischen Kurzzusammenfassungen ermöglichte eine Abschätzung des Einflusses der verschiedenen Verfahrensschritte auf Präzision und Vollständigkeit der Zuordnungen. Die ermittelte Präzision der Zuordnungen lag bei 94 % bei einer Vollständigkeit von 39 %, womit die Qualitätsanforderungen von BRENDA voll erfüllt werden konnten.

Zur Unterstützung der manuellen Inspektion der automatischen Annotation und zur Generierung von Hypothesen für bisher nicht untersuchte Zusammenhänge wurden verschiedene Visualisierungsmöglichkeiten in das Programmpaket eingebaut. Im Zusammenspiel mit den vorgestellten Methoden zur Informationsbeschaffung und der Hypothesenerstellung deckt das in dieser Arbeit entwickelte Programmpaket damit alle Aspekte des *Text Mining* ab.

Als nach Kenntnis des Authors erste Implementierung eines auf Konzepten und semantischen Feldern basierenden Algorithmus für die Extraktion funktioneller Zusammenhänge ist diese Arbeit ein idealer Ausgangspunkt für weitergehende Ansätze, beispielsweise der Entwicklung einer komplexeren Bewertungsfunktion. Mit dem abzusehenden Wachstum von Datenbanken und Volltextarchiven, sowie der dargestellten zunehmenden Präzision der Verfahren zur Informationsextraktion sollten automatische Annotationen zunehmend Akzeptanz finden, um so die große Lücke zwischen beschriebenen und in Datenbanken verfügbaren Zusammenhängen zu schließen.

A ANHANG

A.1 Liste der verwendeten statischen Stoppwörter

a	any	beside	done	first
about	anyhow	besides	down	for
above	anyone	between	due	former
abs	anything	beyond	during	formerly
accordingly	anywhere	both	each	found
across	applicable	but	ec	from
after	apply	by	ed	further
afterwards	are	came	effected	gave
again	arise	can	eg	get
against	around	cannot	either	give
all	as	cc	else	go
almost	assume	cm	elsewhere	gone
alone	at	co	enough	got
along	be	come	especially	gov
already	became	compare	et	had
also	because	could	etc	has
although	become	de	even	have
always	becomes	dealing	ever	having
am	becoming	department	every	he
among	been	depend	everyone	hence
amongst	before	did	everything	her
an	beforehand	discover	everywhere	here
analyze	behind	dl	except	hereafter
and	being	do	few	hereby
another	below	does	find	herein

hereupon
hers
herself
him
himself
his
how
however
hr
i
ie
if
ii
iii
immediately
importance
important
in
inc
incl
indeed
into
investigate
is
it
its
itself
just
keep
kept
kg
km
last
latter
latterly
lb
ld
least
less
letter
like
ltd
made
mainly
make
many
may
me
meanwhile
mg
might
ml
mm
mo
more
moreover
most
mostly
mr
much
mug
must
my
myself
namely
nearly
necessarily
neither
never
nevertheless
next
no
nobody
none
noone
nor
normally
nos
not
noted
nothing
now
nowhere
obtained
of
off
often
on
once
one
only
onto
or
other
others
otherwise
ought
our
ours
ourselves
out
over
overall
owing
own
oz
particularly
per
perhaps
pm
precede
predominantly
present
presently
previously
primarily
promptly
pt
quickly
quite
rather
readily
really
recently
refs
regarding
relate
said
same
seem
seemed
seeming
seems
seen
seriously
several
shall
she
should
show
showed
shown
shows
significantly
since
slightly
so
some
somehow
someone
something
sometime
sometimes
somewhat
somewhere
soon
specifically
still
strongly
studied
sub
substantially
such
sufficiently
take
tell
th
than
that
the
their
theirs
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
they
this
thorough
those
though
through
throughout
thru
thus
to
together
too
toward
towards
try
type
ug
under
unless
until
up
upon
us
use
used
usefully
usefulness
using
usually
various
very
via
was
we
well
were
what

whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
with
within
without
wk
would
wt
yet
you
your
yours
yourself
yourselves
yr

A.2 Krankheitsrelevante MESH-Kategorien

Tabelle A.1 · Liste der krankheitsrelevanten MESH-Kategorien.

MESH-Code	Bezeichnung
C01	Bacterial Infections and Mycoses
C02	Virus Diseases
C03	Parasitic Diseases
C04	Neoplasms
C05	Musculoskeletal Diseases
C06	Digestive System Diseases
C07	Stomatognathic Diseases
C08	Respiratory Tract Diseases
C09	Otorhinolaryngologic Diseases
C10	Nervous System Diseases
C11	Eye Diseases
C12	Urologic and Male Genital Diseases
C13	Female Genital Diseases and Pregnancy Complications
C14	Cardiovascular Diseases
C15	Hemic and Lymphatic Diseases
C16	Neonatal Diseases and Abnormalities
C17	Skin and Connective Tissue Diseases
C18	Nutritional and Metabolic Diseases
C19	Endocrine Diseases
C20	Immunologic Diseases
C21	Disorders of Environmental Origin

A.3 Liste der semantischen Felder in der UMLS-Ontologie

Acquired Abnormality
Activity
Age Group
Alga
Amino Acid Sequence
Amino Acid, Peptide, or Protein
Amphibian
Anatomical Abnormality
Anatomical Structure
Animal
Antibiotic
Archaeon
Bacterium
Behavior
Biologic Function
Biologically Active Substance
Biomedical Occupation or Discipline
Biomedical or Dental Material
Bird
Body Location or Region
Body Part, Organ, or Organ Component
Body Space or Junction
Body Substance
Body System
Carbohydrate Sequence
Carbohydrate
Cell Component
Cell Function
Cell or Molecular Dysfunction
Cell
Chemical Viewed Functionally
Chemical Viewed Structurally
Chemical
Classification
Clinical Attribute
Clinical Drug
Conceptual Entity
Congenital Abnormality
Daily or Recreational Activity
Diagnostic Procedure
Disease or Syndrome
Drug Delivery Device
Educational Activity
Eicosanoid
Element, Ion, or Isotope
Embryonic Structure
Entity
Environmental Effect of Humans
Enzyme
Event
Experimental Model of Disease
Family Group
Finding
Fish
Food
Fully Formed Anatomical Structure
Functional Concept
Fungus
Gene or Genome
Genetic Function
Geographic Area
Governmental or Regulatory Activity
Group Attribute
Group
Hazardous or Poisonous Substance
Health Care Activity
Health Care Related Organization
Hormone
Human
Human-caused Phenomenon or Process
Idea or Concept
Immunologic Factor
Indicator, Reagent, or Diagnostic Aid
Individual Behavior
Injury or Poisoning
Inorganic Chemical
Intellectual Product
Invertebrate
Laboratory or Test Result
Laboratory Procedure

Language
Lipid
Machine Activity
Mammal
Manufactured Object
Medical Device
Mental or Behavioral Dysfunction
Mental Process
Molecular Biology Research Technique
Molecular Function
Molecular Sequence
Natural Phenomenon or Process
Neoplastic Process
Neuroreactive Substance or Biogenic Amine
Nucleic Acid, Nucleoside, or Nucleotide
Nucleotide Sequence
Occupation or Discipline
Occupational Activity
Organ or Tissue Function
Organic Chemical
Organism Attribute
Organism Function
Organism
Organization
Organophosphorus Compound
Pathologic Function
Patient or Disabled Group
Pharmacologic Substance
Phenomenon or Process
Physical Object
Physiologic Function
Plant
Population Group
Professional or Occupational Group
Professional Society
Qualitative Concept
Quantitative Concept
Receptor
Regulation or Law
Reptile
Research Activity
Research Device
Rickettsia or Chlamydia
Self-help or Relief Organization
Sign or Symptom
Social Behavior
Spatial Concept
Steroid
Substance
Temporal Concept
Therapeutic or Preventive Procedure
Tissue
Vertebrate
Virus
Vitamin

A.4 Übersicht der Enzymklassen und Krankheitskonzepte

Ein Auszug der in dieser Arbeit erstellten Zuordnungen von 1 409 Krankheitskonzepten zu 524 Enzymklassen der BRENDA-Datenbank. Eine vollständige Liste findet sich auf dem beiliegenden Datenträger (siehe Abschnitt A.8 auf Seite 139).

Tabelle A.2 · Übersicht der 50 Enzymklassen mit der höchsten Zahl an Krankheitszuordnungen.

EC-Nummer	Empfohlener Name	Anzahl zugeordneter Krankheitskonzepte
3.4.15.1	Peptidyl-dipeptidase A	121
3.4.23.15	Renin	104
3.4.21.36	Pancreatic elastase	97
1.14.13.39	Nitric-oxide synthase	92
3.4.21.5	Thrombin	84
3.4.21.37	Leucocyte elastase	77
3.4.24.3	Microbial collagenase	70
3.1.1.3	Triacylglycerol lipase	67
3.5.4.4	Adenosine deaminase	66
1.15.1.1	Superoxide dismutase	64
3.1.1.7	Acetylcholinesterase	62
3.4.21.4	Trypsin	62
1.16.3.1	Ferroxidase	60
3.6.1.3	Adenosinetriphosphatase	57
1.11.1.9	Glutathione peroxidase	56
3.1.1.34	Lipoprotein lipase	54
3.1.3.1	Alkaline phosphatase	53
3.4.21.69	Protein C (activated)	52
1.11.1.6	Catalase	51
2.7.3.2	Creatine kinase	49
3.2.1.23	beta-Galactosidase	49
3.4.24.7	Interstitial collagenase	47
3.4.21.32	Brachyurin	47

Tabelle A.2 · Fortsetzung

EC-Nummer	Empfohlener Name	Anzahl zugeordneter Krankheitskonzepte
1.1.1.27	L-Lactate dehydrogenase	46
1.11.1.7	Peroxidase	45
3.1.1.4	Phospholipase A2	44
1.9.3.1	Cytochrome-c oxidase	44
3.4.24.23	Matrilysin	44
4.6.1.1	Adenylate cyclase	43
3.4.21.59	Tryptase	42
3.1.1.8	Cholinesterase	40
3.4.21.7	Plasmin	40
3.4.21.68	t-Plasminogen activator	38
3.2.1.18	exo-alpha-Sialidase	37
2.3.1.43	Phosphatidylcholine-sterol O-acyltransferase	37
3.2.1.52	beta-N-acetylhexosaminidase	36
2.3.2.2	gamma-Glutamyltransferase	35
2.7.1.40	Pyruvate kinase	34
3.2.1.31	beta-Glucuronidase	34
1.14.99.10	Steroid 21-monooxygenase	33
3.4.21.73	u-Plasminogen activator	32
4.2.1.1	Carbonate dehydratase	31
4.2.1.11	Phosphoenolpyruvate hydratase	31
3.4.23.35	Barrierpepsin	31
3.2.1.20	alpha-Glucosidase	31
3.5.2.6	beta-Lactamase	30
2.7.1.37	Protein kinase	30
1.4.3.4	Amine oxidase (flavin-containing)	30
3.1.1.47	1-Alkyl-2-acetylglycerophosphocholine esterase	29
1.2.4.1	Pyruvate dehydrogenase (lipoamide)	28

Tabelle A.3 · Übersicht der 50 Krankheitskonzepte mit der höchsten Zahl an zugeordneten Enzymklassen

Konzeptnummer	Konzeptbezeichnung	Anzahl zugeordneter Enzymklassen
C0003873	Rheumatoid Arthritis	100
C0011860	Diabetes Mellitus, Non-Insulin-Dependent	77
C0022661	Kidney Failure, Chronic	51
C0011854	Diabetes Mellitus, Insulin-Dependent	49
C0024141	Lupus Erythematosus, Systemic	45
C0022658	Kidney Diseases	43
C0004153	Atherosclerosis	38
C0155626	Acute myocardial infarction	36
C0004096	Asthma	35
C0035078	Kidney Failure	33
C0026848	Myopathy	33
C0019247	Hereditary Diseases	33
C0001339	Acute pancreatitis	32
C0023890	Liver Cirrhosis	31
C0030567	Parkinson Disease	31
C0011389	Dental Plaque	29
C0027765	Nervous System Diseases	27
C0020179	Huntington Disease	27
C0009373	Colonic Diseases	27
C0036690	Septicemia	27
C0851162	Infections of musculoskeletal system	26
C0006279	Bronchoalveolar Lavage Fluid	26
C0151744	Myocardial Ischemia	25
C0002736	Amyotrophic Lateral Sclerosis	24
C0021390	Inflammatory Bowel Diseases	24
C0009324	Colitis, Ulcerative	24
C0038454	Cerebrovascular accident	24
C0520463	Hepatitis, Chronic Active	23
C0085078	Lysosomal Storage Diseases	23
C0003864	Arthritis	22

Tabelle A.3 · Fortsetzung

Konzeptnummer	Konzeptbezeichnung	Anzahl zugeordneter Enzymklassen
C0033860	Psoriasis	22
C0018801	Heart failure, NOS	22
C0033575	Prostatic Diseases	21
C0014060	Encephalitis, St. Louis	21
C0001175	Acquired Immunodeficiency Syndrome	20
C0017658	Glomerulonephritis	20
C0494792	Primary biliary cirrhosis	20
C0023896	Liver Diseases, Alcoholic	20
C0026769	Multiple Sclerosis	20
C0022660	Kidney Failure, Acute	19
C0012739	Disseminated Intravascular Coagulation	19
C0029408	Osteoarthritis	19
C0025517	Metabolic Diseases	19
C0020550	Hyperthyroidism	19
C0341439	Chronic liver disease	18
C0149521	Chronic pancreatitis, NOS	18
C0010346	Crohn's disease	18
C0034067	Pulmonary Emphysema	18
C0239946	Liver Fibrosis	18

A.5 Karten der KEGG-Datenbank

Tabelle A.4 · Übersicht der Metabolischen Karten der KEGG-Datenbank, die zu einer Gruppierung in sieben Stoffwechselbereiche verwendet wurden (siehe Abschnitt 3.4.2 auf Seite 64).

Gruppierung	Karten der KEGG-Datenbank
Zentralstoffwechsel	Citrate cycle (TCA cycle)
	Fructose and mannose metabolism
	Glycolysis / Gluconeogenesis
	Galactose metabolism
	Pentose and glucuronate interconversions
	Pentose phosphate pathway
	Pyruvate metabolism
Fettsäuren	Fatty acid biosynthesis (path 1)
	Fatty acid biosynthesis (path 2)
	Fatty acid metabolism
	Synthesis and degradation of ketone bodies
Aminosäuren	Alanine and aspartate metabolism
	Aminosugars metabolism
	Arginine and proline metabolism
	beta-Alanine metabolism
	Cysteine metabolism
	Glutamate metabolism
	Glycine, serine and threonine metabolism
	Histidine metabolism
	Lysine biosynthesis
	Lysine degradation
	Phenylalanine metabolism
	Phenylalanine, tyrosine and tryptophan biosynthesis
	Tryptophan metabolism
	Tyrosine metabolism
	Valine, leucine and isoleucine biosynthesis
	Valine, leucine and isoleucine degradation
	Urea cycle and metabolism of amino groups
Nukleotide	Nucleotide sugars metabolism

Tabelle A.4 · Fortsetzung

Gruppierung	Karten der KEGG-Datenbank
	Purine metabolism
	Pyrimidine metabolism
Lipide	Glycerolipid metabolism
	Phospholipid degradation
	Prostaglandin and leukotriene metabolism
	Sphingoglycolipid metabolism
	Sphingophospholipid biosynthesis
Steroide	Androgen and estrogen metabolism
	Bile acid biosynthesis
	C21-Steroid hormone metabolism
	Sterol biosynthesis
Kofaktoren & Vitamine	Biotin metabolism
	Folate biosynthesis
	Pantothenate and CoA biosynthesis
	Porphyrin and chlorophyll metabolism
	Retinol metabolism
	Riboflavin metabolism
	Thiamine metabolism
	Vitamin B6 metabolism

A.6 Datenbankschema

Die aus öffentlichen Datenbanken abgerufenen Einträge und durch die verschiedenen Prozessierungsschritte entstandenen Daten wurden lokal in einer relationalen Datenbank abgelegt. Das Datenbankschema lässt sich grob in sechs Bereiche unterteilen, die in Abbildung A.1 entsprechend farblich hervorgehoben sind. Die Tabellen sind (*kursiv*) und mit Klammerung angegeben:

- BRENDA: Die BRENDA-Datenbank bildet die Grundlage zur Erstellung des Enzymlexikons aus den empfohlenen Enzymnamen (*Recommended Names*) sowie deren Synonymen (*Synonyms*).
- *PubMed*: Die über *PubMed* erhaltenen Dokumente befinden sich zusammen mit ihrer Identifikationsnummer in der Tabelle (*PubMed*). Nach der Extraktion von Titel und Kurzzusammenfassung sowie den von Annotatoren zugewiesenen Schlagwörtern erfolgt die Speicherung in den Tabellen (*Document*, *MeshLink* und *Mesh*).
- Lexikalische Analyse: Die nach der lexikalischen Prozessierung vorliegenden Informationen finden sich in verschiedenen Tabellen wieder. Die Tabellen (*WordLink*) und (*Word*) enthalten Informationen über gefundene Wörter sowie deren Position im Text; für Stammformen sind die Tabellen (*Stem*) und (*StemLink*) zuständig. Die über das *MetaMap*-Programmpaket in Kurzzusammenfassungen identifizierten Konzepte der UMLS-Ontologie sind (zusammen mit ihrer Identifikationsnummer – CUI für *concept unique identifier*) in den Tabellen (*ConceptLink*) und (*MeshCUI*) gespeichert. Die mit Hilfe des Enzymlexikons identifizierten Enzymnamen befinden sich in Tabelle (*ECLink*).
- UMLS: Eine Integration mehrerer Tabellen der UMLS-Ontologie in die Datenbank erleichtert die Analyse der Konzeptzuordnung, diese sind im Schema als UMLS gruppiert:
  - Semantische Felder: Jeder Konzeptnummer CUI sind ein oder mehrere semantische Felder zugewiesen (*MRSTY*).

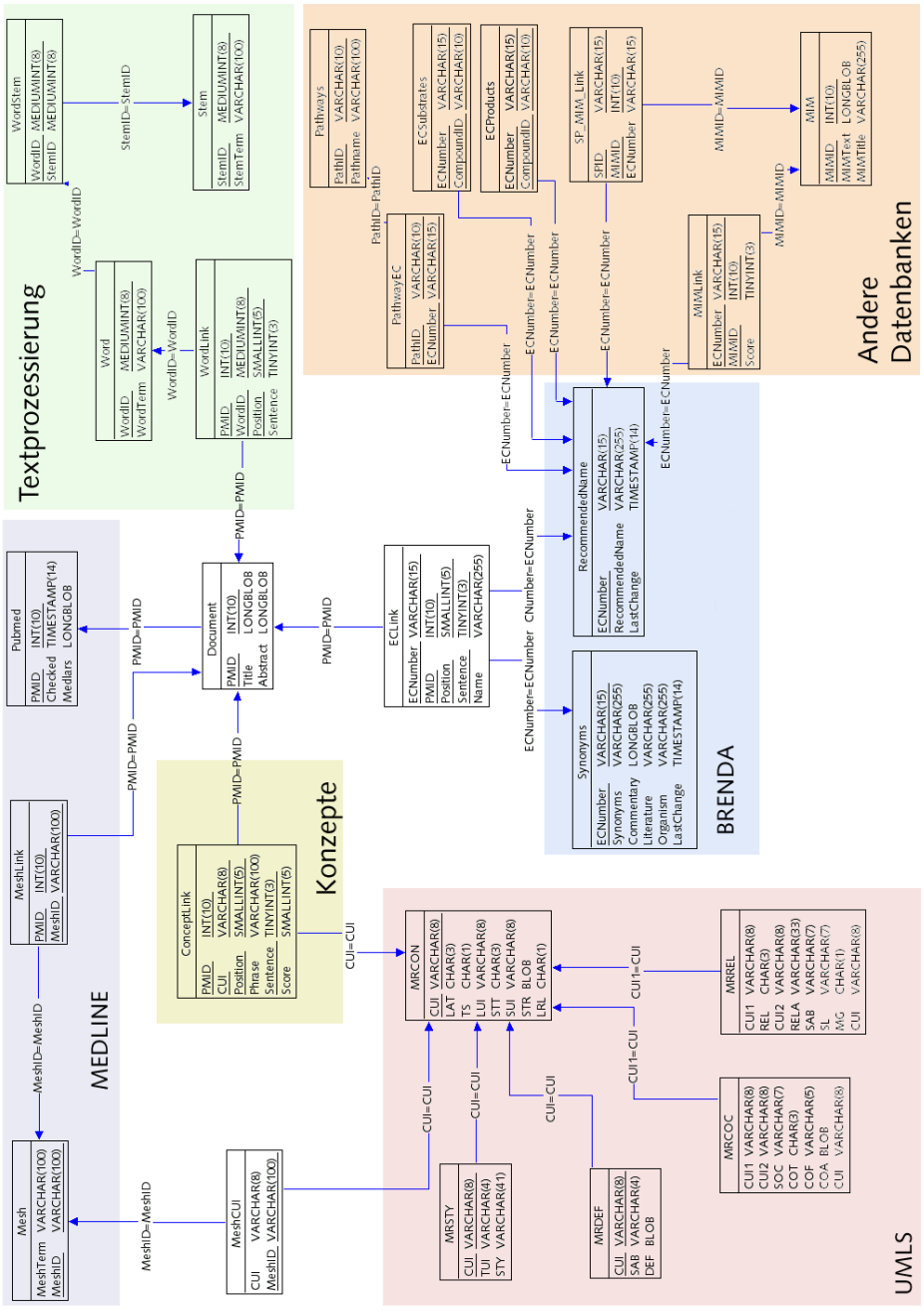


Abbildung A.1 · Schema der Datenbank. Die Tabellen sind je nach Datenquelle oder Prozessierungsschritt farblich zusammengefasst, eine kurze Beschreibung der einzelnen Tabellen und Zusammenhänge findet sich im Text.

  - Konzeptbeschreibungen: Jedes Konzept besitzt Verweise auf seine Quelle sowie ein oder mehrere Namen (*MRCON*). Eine eindeutige Definition jedes Konzepts findet sich in Tabelle (*MRDEF*).
  - Beziehungen zwischen Konzepten: Neben dem semantischen Netzwerk, das die Beziehungen jedes Konzepts zu anderen verwandten Konzepten wiedergibt (*MRREL*), enthält die UMLS-Ontologie auch Informationen zu besonders häufig in den annotierten *PubMed*-Dokumenten gemeinsam auftretenden Konzepten (*MRCOC*).

- Andere Datenbanken: Die Einträge der OMIM-Datenbank sind in Tabelle (*MIM*), ihre Zuordnung zu Enzymklassen in Tabelle (*MIM-Link*) gespeichert. Die aus *Swiss-Prot*-Einträgen extrahierten Verweise auf EC-Nummern finden sich in Tabelle (*SP_EC_Link*), die Verweise auf OMIM in Tabelle (*SP_MIM_Link*). In der KEGG-Datenbank finden sich Informationen über die Beteiligung von Enzymen verschiedener Enzymklassen an unterschiedlichen Stoffwechselwegen (*PathwayEC, Pathways*). Ebenfalls aus der KEGG-Datenbank stammen die Daten zu den Substrat-Produkt-Ketten der Enzymklassen (*ECProducts, ECSubstrates*).

A.7 Verwendete Software

Verwendete Programme:

Python 2.2	http://python.org/
MySQL 3.23.45	http://www.mysql.com/
MetaMap 2.2c	http://mmtx.nlm.nih.gov/
libSVM 2.5	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
Touchgraph LB 1.2	http://touchgraph.sourceforge.net/
Walrus 0.6	http://www.caida.org/tools/visualization/walrus/
Pajek 0.97	http://vlado.fmf.uni-lj.si/pub/networks/pajek/
Dislin 8.1	http://www.linmpi.mpg.de/dislin/

Zusätzliche Module für Python:

Numeric Python 22.0	http://www.pfdubois.com/numpy/
BioPython 1.22	http://biopython.org/
orange 0.9	http://magix.fri.uni-lj.si/orange/
pExpect 0.9	http://pexpect.sourceforge.net/
mySQLdb 0.9	http://sourceforge.net/projects/mysql-python
Disipyl 0.8	http://kim.bio.upenn.edu/~pmagwene/

A.8 Inhalt der CD-ROM

Annotation/	Annotierte Textkorpora
Database/	Inhalt der aufgebauten Datenbank im MySQL-Format
KEGG/	Metabolische Pfadanalysen
Gaps/	Informationen zu Lücken in der Annotation
KEGG_Images/	Abbildungen der Pfade
KEGG_Input/	Steuerdateien für *tagKEGGMap*
Touchgraph_Input	Steuerdateien für *TouchGraph*
Matrices/	Textähnlichkeit für Krankheiten
OMIM/	Zuordnungen von Enzymklassen zu OMIM-Einträgen
Otter/	Steuerdateien für die Visualisierung mit *Otter*
Pajek/	Daten zur Netzwerkanalyse mit *Pajek*
Paper/	Publikationen und Poster zur Dissertation
Presentations/	Vortragsunterlagen
Reports/	Zwischenberichte
Sequences/	Vergleich der Krankheitszuordnungen mit Enzymsequenzen
Source/	Quellen und Dokumentation des entwickelten Programms
Thesis/	Dissertation mit LaTeX-Quellen und im PDF-Format
Touchgraph/	Steuerdateien für die Visualisierung mit *Touchgraph*
TreeView/	Gruppierungen der Enzymklassen mit *Cluster*
Walrus/	Steuerdateien für die Visualisierung mit *Walrus*
README	Informationen zu den enthaltenen Daten

A.9 Vorabveröffentlichungen

Ida Schomburg, Oliver Hofmann, Claudia Bänsch, Antje Chang, Dietmar Schomburg (2000): Enzyme data and metabolic information: brenda, a resource for research in biology, biochemistry, and medicine. *Gene Function & Disease, 1, (3-4):109*

Ida Schomburg, Antje Chang, Oliver Hofmann, Christian Ebeling, Frank Ehrentreich, Dietmar Schomburg (2002): brenda: a resource for enzyme data and metabolic information. *Trends Biochem Sci. 27(1):54-56. Review*

Oliver Hofmann, Dietmar Schomburg (2002): Mining the literature for disease-enzyme associations. *ISMB Poster Präsentation, Edmonton, Kanada*

Oliver Hofmann, Dietmar Schomburg (2002): Enzyme-disease networks extracted from the scientific literature. *ECCB Poster Präsentation, Saarbrücken*

Oliver Hofmann, Dietmar Schomburg (2003): Enzyme-disease associations extracted from the biomedical literature. *GCB Poster Präsentation, München*

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Allen, James: *Natural Language Understanding*. 2nd. Pearson Addison Wesley, 1994

[2] Andrade, M. A. ; Bork, P.: Automated extraction of information in molecular biology. In: *FEBS Lett* 476 (2000), Nr. 1-2, S. 12–7

[3] Andrade, M. A. ; Valencia, A.: Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. In: *Proc Int Conf Intell Syst Mol Biol* 5 (1997), S. 25–32

[4] Andrade, M. A. ; Valencia, A.: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. In: *Bioinformatics* 14 (1998), Nr. 7, S. 600–7

[5] Aronson, A. R. ; Rindflesch, T. C.: Query expansion using the UMLS Metathesaurus. In: *Proc AMIA Annu Fall Symp* (1997), S. 485–9

[6] Aronson, Alan R.: MetaMap Candidate Retrieval / Semantic Knowledge Representation Group. 2001. – Forschungsbericht

[7] Aronson, Alan R.: MetaMap Mapping Algorithm / Semantic Knowledge Representation Group. 2001. – Forschungsbericht

[8] Aronson, Alan R.: MetaMap Variant Generation / Semantic Knowledge Representation Group. 2001. – Forschungsbericht

[9] Aronson, Alan R.: Filtering the UMLS Metathesaurus for MetaMap / Semantic Knowledge Representation Group. 2002. – Forschungsbericht

[10] Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: *Proc AMIA Symp* (2001), S. 17–21

[11] Babbitt, P. C.: Definitions of enzyme function for the structural genomics era. In: *Curr Opin Chem Biol* 7 (2003), Nr. 2, S. 230–7

[12] Baclawski, K. ; Cigna, J. ; Kokar, M. M. ; Mager, P. ; Indurkhya, B.: Knowledge representation and indexing using the unified medical language system. In: *Pac Symp Biocomput* (2000), S. 493–504

[13] Barabasi, A. L. ; Albert, R.: Emergence of scaling in random networks. In: *Science* 286 (1999), Nr. 5439, S. 509–12

[14] Barabasi, Albert-Laszlo ; Albert, Reka ; Hawoong, Jeong: Scale-free characteristics of random networks: The topology of the World Wide Web. In: *Physica A*

(2000), Nr. 281, S. 67–77

[15] Batagelj, V ; Mrvar, A.: Pajek – Program for Large Network Analysis. In: *Connections* 21 (1998), Nr. 2

[16] Blaschke, C. ; Andrade, M. A. ; Ouzounis, C. ; Valencia, A.: Automatic extraction of biological information from scientific text: protein-protein interactions. In: *Proc Int Conf Intell Syst Mol Biol* (1999), S. 60–7

[17] Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. In: *Nucleic Acids Res* 32 Database issue (2004), S. D267–70

[18] Bodenreider, O. ; Burgun, A. ; Rindflesch, T. C.: Assessing the consistency of a biomedical terminology through lexical knowledge. In: *Int J Med Inf* 67 (2002), Nr. 1-3, S. 85–95

[19] Bodenreider, O. ; Nelson, S.J. ; Hole, W.T. ; Chang, H.F.: Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In: *Proc AMIA Symp* (1998), S. 815–9

[20] Bodenreider, Olivier ; Rindflesch, Thomas ; Burgun, Anita: Unsupervised, corpus-based method for extending a biomedical terminology. In: *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia : Association for Computational Linguistics, July 2002, S. 53–60

[21] Boeckmann, B. ; Bairoch, A. ; Apweiler, R. ; Blatter, M. C. ; Estreicher, A. ; Gasteiger, E. ; Martin, M. J. ; Michoud, K. ; O'Donovan, C. ; Phan, I. ; Pilbout, S. ; Schneider, M.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. In: *Nucleic Acids Res* 31 (2003), Nr. 1, S. 365–70

[22] Bomalaski, J. S. ; Clark, M. A. ; Zurier, R. B.: Enhanced phospholipase activity in peripheral blood monocytes from patients with rheumatoid arthritis. In: *Arthritis Rheum* 29 (1986), Nr. 3, S. 312–8

[23] Bonafe, L. ; Blau, N. ; Burlina, A. P. ; Romstad, A. ; Guttler, F. ; Burlina, A. B.: Treatable neurotransmitter deficiency in mild phenylketonuria. In: *Neurology* 57 (2001), Nr. 5, S. 908–11

[24] Boyle, M. D. ; Ohanian, S. H. ; Borsos, T.: Lysis of tumor cells by antibody and complement. VI. Enhanced killing of enzyme-pretreated tumor cells. In: *J Immunol* 116 (1976), Nr. 3, S. 661–8

[25] Bruijn, B. de ; Martin, J.: Getting to the (c)ore of knowledge: mining biomedical literature. In: *Int J Med Inf* 67 (2002), Nr. 1-3, S. 7–18

[26] Bruijn, Berry de ; Martin, Joel: Literature Mining in Molecular Biology. In: *Proceedings of the EFMI workshop on Natural Language Processing in Biomedical Applications*. Nicosia, Cyprus : R. Baud and P. Ruch, March 2002, S. 1–5

[27] Burlina, A. B. ; Bonafe, L. ; Ferrari, V. ; Suppiej, A. ; Zacchello, F. ; Bur-

LINA, A. P.: Measurement of neurotransmitter metabolites in the cerebrospinal fluid of phenylketonuric patients under dietary treatment. In: *J Inherit Metab Dis* 23 (2000), Nr. 4, S. 313–6

[28] BUSH, Vannevar: As we may think. In: *Atlantic Monthly* 176 (1945), July, Nr. 1, S. 101–108

[29] CABEZAS, Clara ; RESNIK, Philip ; STEVENS, Jessica: Supervised Sense Tagging using Support Vector Machines. In: *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 2001

[30] CAMON, E. ; MAGRANE, M. ; BARRELL, D. ; LEE, V. ; DIMMER, E. ; MASLEN, J. ; BINNS, D. ; HARTE, N. ; LOPEZ, R. ; APWEILER, R.: The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. In: *Nucleic Acids Res* 32 Database issue (2004), S. D262–6

[31] CAMPBELL, S. J. ; GOLD, N. D. ; JACKSON, R. M. ; WESTHEAD, D. R.: Ligand binding: functional site location, similarity and docking. In: *Curr Opin Struct Biol* 13 (2003), Nr. 3, S. 389–95

[32] CHABROL, B. ; MANCINI, J. ; BENELLI, C. ; GIRE, C. ; MUNNICH, A.: Leigh syndrome: pyruvate dehydrogenase defect. A case with peripheral neuropathy. In: *J Child Neurol* 9 (1994), Nr. 1, S. 52–5

[33] CHACKO, C.M. ; CHRISTIAN, J.C. ; NADLER, H.L.: Unstable galactose-1-phosphate uridyl transferase: a new variant of galactosemia. In: *J Pediatr* (1971)

[34] CHANG, Chih-Chung ; LIN, Chih-Jen: libSVM – A Library for Support Vector Machines / Computer Science and Information Engineering, National Taiwan University. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2004. – Forschungsbericht

[35] CHANG, J. T. ; SCHUTZE, H. ; ALTMAN, R. B.: GAPSCORE: finding gene and protein names one word at a time. In: *Bioinformatics* 20 (2004), Nr. 2, S. 216–25

[36] CHANG, J.T. ; RAYCHAUDHURI, S. ; ALTMAN, R.B.: Including biological literature improves homology search. In: *Pac Symp Biocomput* (2001), S. 374–83

[37] CHISHOLM, Erica ; KOLDA, Tamara G.: New term weighting formulas for the vector space method in information retrieval / Oak Ridge National Laboratory. Oak Ridge, Tenessee, 1988 (ORNL/TM-13756). – Forschungsbericht

[38] CHOMSKY, N.: *Syntactic Structures*. The Hague:Mouton, 1957

[39] COHEN, K. B. ; DOLBEY, Andrew ; ACQUAAH-MENSAH, George ; HUNTER, Lawrence: Contrast and variability in gene names. In: *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia : Association for Computational Linguistics, July 2002, S. 14–20

[40] COUTINHO, P. M. ; HENRISSAT, B.: *Carbohydrate-Active Enzymes server*. Data-

base. – URL http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html

[41] Craven, M. ; Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources. In: *Proc Int Conf Intell Syst Mol Biol* (1999), S. 77–86

[42] Craven, Mark: Learning to extract relations from MEDLINE. In: *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999

[43] Craven, Mark: Hierarchical Hidden Markov Models for Information Extraction. In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence* (2003)

[44] Cutting, Doug ; Kupiec, Julian ; Pedersen, Jan ; Sibun, Penelope: Practical Part-of-Speech Tagger. In: *Proceedings of ANLP*, 1992

[45] Davidsen, J. ; Ebel, H. ; Bornholdt, S.: Emergence of a small world from local interactions: modelling acquaintance networks. In: *Phys. Rev. Lett.* (2002), Nr. 88

[46] Demetriou, George ; Gaizauskas, Robert: Utilizing text mining results: The Pasta Web System. In: *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia : Association for Computational Linguistics, July 2002, S. 77–84

[47] Dhein, S. ; Hochreuther, S. ; Aus Dem Spring, C. ; Bollig, K. ; Hufnagel, C. ; Raschack, M.: Long-term effects of the endothelin(A) receptor antagonist LU 135252 and the angiotensin-converting enzyme inhibitor trandolapril on diabetic angiopathy and nephropathy in a chronic type I diabetes mellitus rat model. In: *J Pharmacol Exp Ther* 293 (2000), Nr. 2, S. 351–9

[48] Dickman, S.: Tough Mining: The challenges of searching the scientific literature. In: *PLoS Biol* 1 (2003), Nr. 2, S. E48

[49] DiLella, A. G. ; Marvit, J. ; Brayton, K. ; Woo, S. L.: An amino-acid substitution involved in phenylketonuria is in linkage disequilibrium with DNA haplotype 2. In: *Nature* 327 (1987), Nr. 6120, S. 333–6

[50] Ding, J. ; Berleant, D. ; Nettleton, D. ; Wurtele, E.: Mining MEDLINE: abstracts, sentences, or phrases? In: *Pac Symp Biocomput* (2002), S. 326–37

[51] Dobrokhotov, P. B. ; Goutte, C. ; Veuthey, A. L. ; Gaussier, E.: Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. In: *Bioinformatics* 19 Suppl 1 (2003), S. i91–4

[52] Doekes, G. ; Kamminga, N. ; Helwegen, L. ; Heederik, D.: Occupational IgE sensitisation to phytase, a phosphatase derived from Aspergillus niger. In: *Occup Environ Med* 56 (1999), Nr. 7, S. 454–9

[53] Donaldson, I. ; Martin, J. ; Bruijn, B. de ; Wolting, C. ; Lay, V. ; Tuekam, B. ; Zhang, S. ; Baskin, B. ; Bader, G. D. ; Michalickova, K. ; Pawson,

T. ; Hogue, C. W.: PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. In: *BMC Bioinformatics* 4 (2003), Nr. 1, S. 11

[54] Eibe, Frank ; Paynter, Gordon W. ; Witten, Ian H.: Domain-Specific Keyphrase Extraction. In: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999, S. 668–673

[55] Eisen, M.B. ; Spellman, P.T. ; Brown, P.O. ; Botstein, D.: Cluster analysis and display of genome-wide expression patterns. In: *Proc Natl Acad Sci* 95 (1998), Nr. 25, S. 14863–8

[56] Erdös, P. ; A., Réyni.: Random Graphs. In: *Publ Math Inst Hung Acad Sci* (1960)

[57] Faraj, B. A. ; Newman, S. L. ; Caplan, D. B. ; Ahmann, P. A. ; Kutner, M. ; Ali, F. M. ; Lindahl, J. A.: Platelet-monoamine oxidase activity in Reye's syndrome. In: *J Pediatr Gastroenterol Nutr* 4 (1985), Nr. 4, S. 532–6

[58] Feldman, Ronen ; Regev, Yizhar ; Hurvitz, Eyal ; Finkelstein-Landau, Michael: Mining the biomedical literature using semantic analysis of natural language processing techniques. In: *Biosilico* 1 (2003), Nr. 2

[59] Fell, D. A. ; Wagner, A.: The small world of metabolism. In: *Nat Biotechnol* 18 (2000), Nr. 11, S. 1121–2

[60] Fields, S. ; Song, O.K.: A novel genetic system to detect protein-protein interactions. In: *Nature* 340 (1989)

[61] Forster, M. J.: Molecular modelling in structural biology. In: *Micron* 33 (2002), Nr. 4, S. 365–84

[62] Fridovich-Keil, J. L. ; Jinks-Robertson, S.: A yeast expression system for human galactose-1-phosphate uridylyltransferase. In: *Proc Natl Acad Sci U S A* 90 (1993), Nr. 2, S. 398–402

[63] Gaizauskas, R. ; Demetriou, G. ; Artymiuk, P. J. ; Willett, P.: Protein structures and information extraction from biological texts: the PASTA system. In: *Bioinformatics* 19 (2003), Nr. 1, S. 135–43

[64] Galperin, M. Y.: The Molecular Biology Database Collection: 2004 update. In: *Nucleic Acids Res* 32 Database issue (2004), S. D3–22

[65] Guare, John: *Six Degrees of Seperation*. Random House, 1990

[66] Hahn, U. ; Romacker, M. ; Schulz, S.: Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. In: *Pac Symp Biocomput* (2002), S. 338–49

[67] Hanisch, D. ; Fluck, J. ; Mevissen, H. T. ; Zimmer, R.: Playing biology's name game: identifying protein names in scientific text. In: *Pac Symp Biocomput* (2003), S. 403–14

[68] Harris, M. A. ; Clark, J. ; Ireland, A. ; Lomax, J. ; Ashburner, M. ; Foulger, R. ; Eilbeck, K. ; Lewis, S. ; Marshall, B. ; Mungall, C. ; Richter, J. ; Rubin, G. M. ; Blake, J. A. ; Bult, C. ; Dolan, M. ; Drabkin, H. ; Eppig, J. T. ; Hill, D. P. ; Ni, L. ; Ringwald, M. ; Balakrishnan, R. ; Cherry, J. M. ; Christie, K. R. ; Costanzo, M. C. ; Dwight, S. S. ; Engel, S. ; Fisk, D. G. ; Hirschman, J. E. ; Hong, E. L. ; Nash, R. S. ; Sethuraman, A. ; Theesfeld, C. L. ; Botstein, D. ; Dolinski, K. ; Feierbach, B. ; Berardini, T. ; Mundodi, S. ; Rhee, S. Y. ; Apweiler, R. ; Barrell, D. ; Camon, E. ; Dimmer, E. ; Lee, V. ; Chisholm, R. ; Gaudet, P. ; Kibbe, W. ; Kishore, R. ; Schwarz, E. M. ; Sternberg, P. ; Gwinn, M. ; Hannick, L. ; Wortman, J. ; Berriman, M. ; Wood, V. ; Cruz, N. de la ; Tonellato, P. ; Jaiswal, P. ; Seigfried, T. ; White, R.: The Gene Ontology (GO) database and informatics resource. In: *Nucleic Acids Res* 32 Database issue (2004), S. D258–61

[69] Hatzivassiloglou, V. ; Weng, W.: Learning anchor verbs for biological interaction patterns from published text articles. In: *Int J Med Inf* 67 (2002), Nr. 1-3, S. 19–32

[70] Hersh, W. R. ; Hickam, D. H. ; Haynes, R. B. ; McKibbon, K. A.: A performance and failure analysis of SAPHIRE with a MEDLINE test collection. In: *J Am Med Inform Assoc* 1 (1994), Nr. 1, S. 51–60

[71] Hirschman, L. ; Park, J. C. ; Tsujii, J. ; Wong, L. ; Wu, C. H.: Accomplishments and challenges in literature data mining for biology. In: *Bioinformatics* 18 (2002), Nr. 12, S. 1553–61

[72] Huffaker, B. ; Nemeth, E. ; Claffy, K.: Otter: A general-purpose network visualization tool / CAIDA: Cooperative Association for Internet Data Analysis. URL http://www.caida.org/tools/visualization/otter/paper/, 1999. – Forschungsbericht

[73] Humphreys, K. ; Demetriou, G. ; Gaizauskas, R.: Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. In: *Pac Symp Biocomput* (2000), S. 505–16

[74] Huterer, S. ; Wherrett, J. R. ; Poulos, A. ; Callahan, J. W.: Deficiency of phospholipase C acting on phosphatidylglycerol in Niemann-Pick disease. In: *Neurology* 33 (1983), Nr. 1, S. 67–73

[75] Hyun, Young: Walrus – graph visualization tool / Cooperative Association for Internet Data Analysis. University of San Diego, California, USA, 2003. – Forschungsbericht. – URL http://www.caida.org/tools/visualization/walrus

[76] Iliopoulos, I. ; Enright, A. J. ; Ouzounis, C. A.: Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. In: *Pac Symp Biocomput* (2001), S. 384–95

[77] IUBMB: *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology: Classification and Nomenclature of Enzyme-Catalysed*

Reactions. 2002. – URL http://www.chem.qmul.ac.uk/iubmb/enzyme/rules.html

[78] Iwanaga, N. ; Yamamasu, S. ; Tachibana, D. ; Nishio, J. ; Nakai, Y. ; Shintaku, H. ; Ishiko, O.: Activity of synthetic enzymes of tetrahydrobiopterin in the human placenta. In: *Int J Mol Med* 13 (2004), Nr. 1, S. 117–20

[79] Jensen, L. J. ; Gupta, R. ; Blom, N. ; Devos, D. ; Tamames, J. ; Kesmir, C. ; Nielsen, H. ; Staerfeldt, H. H. ; Rapacki, K. ; Workman, C. ; Andersen, C. A. ; Knudsen, S. ; Krogh, A. ; Valencia, A. ; Brunak, S.: Prediction of human protein function from post-translational modifications and localization features. In: *J Mol Biol* 319 (2002), Nr. 5, S. 1257–65

[80] Jenssen, T.K. ; Laegreid, A. ; Komorowski, J. ; Hovig, E.: A literature network of human genes for high-throughput analysis of gene expression. In: *Nat Genet* 28 (2001), Nr. 1, S. 21–8

[81] Jeong, H. ; Mason, S. P. ; Barabasi, A. L. ; Oltvai, Z. N.: Lethality and centrality in protein networks. In: *Nature* 411 (2001), Nr. 6833, S. 41–2

[82] Jeong, H. ; Tombor, B. ; Albert, R. ; Oltvai, Z. N. ; Barabasi, A. L.: The large-scale organization of metabolic networks. In: *Nature* 407 (2000), Nr. 6804, S. 651–4

[83] Kanehisa, M. ; Goto, S. ; Kawashima, S. ; Nakaya, A.: The KEGG databases at GenomeNet. In: *Nucleic Acids Res* 30 (2002), Nr. 1, S. 42–6

[84] Karp, P. D. ; Paley, S. ; Zhu, J.: Database verification studies of SWISS-PROT and GenBank. In: *Bioinformatics* 17 (2001), Nr. 6, S. 526–32; discussion 533–4

[85] Karp, P. D. ; Riley, M. ; Paley, S. M. ; Pellegrini-Toole, A. ; Krummenacker, M.: Eco Cyc: encyclopedia of Escherichia coli genes and metabolism. In: *Nucleic Acids Res* 27 (1999), Nr. 1, S. 55–8

[86] Kazama, Jun'ichi ; Makino, Takaki ; Ohta, Yoshihiro ; Tsujii, Jun'ichi: Tuning support vector machines for biomedical named entity recognition. In: *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia : Association for Computational Linguistics, July 2002, S. 1–8

[87] Kim, J. D. ; Ohta, T. ; Tateisi, Y. ; Tsujii, J.: GENIA corpus–semantically annotated corpus for bio-textmining. In: *Bioinformatics* 19 Suppl 1 (2003), S. i180–2

[88] Kim, Y. B. ; Nikoulina, S. E. ; Ciaraldi, T. P. ; Henry, R. R. ; Kahn, B. B.: Normal insulin-dependent activation of Akt/protein kinase B, with diminished activation of phosphoinositide 3-kinase, in muscle in type 2 diabetes. In: *J Clin Invest* 104 (1999), Nr. 6, S. 733–41

[89] Kostoff, R. N. ; DeMarco, R. A.: Extracting information from the literature by text mining. In: *Anal Chem* 73 (2001), Nr. 13, S. 370A–378A

[90] Krauthammer, M. ; Rzhetsky, A. ; Morozov, P. ; Friedman, C.: Using BLAST for identifying gene and protein names in journal articles. In: *Gene* 259 (2000), Nr. 1-2, S. 245–52

[91] Laaken, E.H. Klein-van der ; Weeber, M. ; Berg, L.T.W. Jong-van den ; Vos, R.: Evaluating MetaMap's text-to-concept mapping performance. (1999)

[92] Langley, Pat: The computer-aided discovery of scientific knowledge. In: *Proceedings of the first international conference on discovery science*, 1998

[93] Lee, Ki-Joong ; Hwang, Young-Sook ; Rim, Hae-Chang: Two-Phase Biomedical NE Recognition based on SVMs. In: Ananiadou, Sophia (Hrsg.) ; Tsujii, Jun'ichi (Hrsg.): *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, URL http://www.aclweb.org/anthology/W03-1305.pdf, 2003, S. 33–40

[94] Lemke, N. ; Heredia, F. ; Barcellos, C. K. ; Dos Reis, A. N. ; Mombach, J. C.: Essentiality and damage in metabolic networks. In: *Bioinformatics* 20 (2004), Nr. 1, S. 115–9

[95] Levenshtein, V.I.: Binary codes capable of correcting deletions insertions and reversals. In: *Soviet Physics-Doklandy* 10 (1966), Nr. 8

[96] Liu, H. ; Johnson, S.B. ; Friedman, C.: Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. In: *J Am Med Inform Assoc* 9 (2002), Nr. 6, S. 621–36

[97] Lord, P. W. ; Stevens, R. D. ; Brass, A. ; Goble, C. A.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. In: *Bioinformatics* 19 (2003), Nr. 10, S. 1275–83

[98] Manning, Christopher D. ; Schütze, Hinrich: *Foundations of Statistical Natural Language Processing*. MIT Press, 1999

[99] Marc, Weeber; ; Klein, Henry ; Loljke, TW ; Berg, Jong van den ; Vos, Rein: Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. In: *JASIST* 52 (2001), Nr. 7, S. 548–577

[100] Marcotte, E. M. ; Xenarios, I. ; Eisenberg, D.: Mining literature for protein-protein interactions. In: *Bioinformatics* 17 (2001), Nr. 4, S. 359–63

[101] McKusick, V.A: Online Mendelian Inheritance in Man OMIM (TM) / McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). URL http://www.ncbi.nlm.nih.gov/omim/, 2000. – Forschungsbericht

[102] Morgan, Alex ; Hirschman, Lynette ; Yeh, Alexander ; Colosimo, Marc: Gene Name Extraction Using FlyBase Resources. In: *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 2003, S. 1–8

[103] Mukhopadhyay, A. ; Deshmukh, D. R. ; Sarnaik, A. P.: Hepatic polyamine metabolism in children with Reye's syndrome. In: *Enzyme* 45 (1991), Nr. 4, S. 209–14

[104] Muller, U. ; Steinberger, D. ; Nemeth, A. H.: Clinical and molecular genetics of primary dystonias. In: *Neurogenetics* 1 (1998), Nr. 3, S. 165–77

[105] Mutalik, P.G. ; Deshpande, A. ; Nadkarni, P.M.: Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. In: *J Am Med Inform Assoc* 8 (2001), Nr. 6, S. 598–609

[106] Nelson, S ; Powerl, T ; Humhphreys, B: The Unified Medical Language System (UMLS) Project / National Library of Medicine. URL http://www.nlm.nih.gov/mesh/umlsforelis.html, 2001. – Forschungsbericht

[107] Nenadic, G. ; Spasic, I. ; Ananiadou, S.: Terminology-driven mining of biomedical literature. In: *Bioinformatics* 19 (2003), Nr. 8, S. 938–43

[108] Newman, M. E.: The structure of scientific collaboration networks. In: *Proc Natl Acad Sci U S A* 98 (2001), Nr. 2, S. 404–9

[109] Ng, S. K. ; Wong, M.: Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. In: *Genome Inform Ser Workshop* 10 (1999), S. 104–112

[110] Niederwieser, A. ; Blau, N. ; Wang, M. ; Joller, P. ; Atares, M. ; Cardesa-Garcia, J.: GTP cyclohydrolase I deficiency, a new enzyme defect causing hyperphenylalaninemia with neopterin, biopterin, dopamine, and serotonin deficiencies and muscular hypotonia. In: *Eur J Pediatr* 141 (1984), Nr. 4, S. 208–14

[111] Ono, T. ; Hishigaki, H. ; Tanigami, A. ; Takagi, T.: Automated extraction of information on protein-protein interactions from the biological literature. In: *Bioinformatics* 17 (2001), Nr. 2, S. 155–61

[112] orange: *A component-based data mining software*. – URL http://magix.fri.uni-lj.si/orange/default.asp

[113] Park, J. C. ; Kim, H. S. ; Kim, J. J.: Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In: *Pac Symp Biocomput* (2001), S. 396–407

[114] Pisanelli, D. M. ; Gangemi, A. ; Steve, G.: An ontological analysis of the UMLS Methathesaurus. In: *Proc AMIA Symp* (1998), S. 810–4

[115] Porter, M.F.: An algorithm for suffix stripping. In: *Programs* 14 (1980), Nr. 3, S. 130–37

[116] Pratt, Meliha: LitLinker: capturing connections across the biomedical literature. In: *Proceedings of the international conference on Knowledge capture*, ACM Press, 2003, S. 105–112. – ISBN 1-58113-583-1

[117] Pustejovsky, J. ; Castano, J. ; Zhang, J. ; Kotecki, M. ; Cochran, B.: Robust

relational parsing over biomedical literature: extracting inhibit relations. In: *Pac Symp Biocomput* (2002), S. 362–73

[118] Reichardt, J. K. ; Woo, S. L.: Molecular basis of galactosemia: mutations and polymorphisms in the gene encoding human galactose-1-phosphate uridylyltransferase. In: *Proc Natl Acad Sci U S A* 88 (1991), Nr. 7, S. 2633–7

[119] Reiter, S. ; Simmonds, H. A. ; Zollner, N. ; Braun, S. L. ; Knedel, M.: Demonstration of a combined deficiency of xanthine oxidase and aldehyde oxidase in xanthinuric patients not forming oxipurinol. In: *Clin Chim Acta* 187 (1990), Nr. 3, S. 221–34

[120] Renner, A. ; Aszodi, A.: High-throughput functional annotation of novel gene products using document clustering. In: *Pac Symp Biocomput* (2000), S. 54–68

[121] Rindflesch, T. C. ; Aronson, A. R.: Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In: *Proc Annu Symp Comput Appl Med Care* (1994), S. 240–4

[122] Rindflesch, T. C. ; Tanabe, L. ; Weinstein, J. N. ; Hunter, L.: EDGAR: extraction of drugs, genes and relations from the biomedical literature. In: *Pac Symp Biocomput* (2000), S. 517–28

[123] Rindflesch, T.C.: Natural Language Processing. In: *Annual Review of Applied Linguistics* 16 (1996), S. 71–85

[124] Rindflesch, Thomas C. ; Hunter, Lawrence ; Aronson, Alan R.: Mining molecular binding terminology from biomedical text. In: *Proc AMIA Symp*, 1999

[125] Rindflesch, Thomas C. ; Rajan, Jayant V. ; Lawrence, Hunter: Extracting Molecular Binding Relationships from Biomedical Text. In: *Applied Natural Language Processing*, 2000

[126] Saksela, M. ; Raivio, K. O.: Cloning and expression in vitro of human xanthine dehydrogenase/oxidase. In: *Biochem J* 315 (Pt 1) (1996), S. 235–9

[127] Salton, G.: Associative document retrieval techniques using bibliographic information. In: *Journal of the ACM* 10 (1963), Nr. 4, S. 440–457

[128] Salton, G. ; Buckley, C.: Term weighting approaches in automatic text retrieval. In: *Information Processing and management* 2 (1989), Nr. 5, S. 513–523

[129] Salton, G. ; Wong, A. ; Yang, C. S.: A vector space model for automatic indexing. In: *Communications of the ACM* 18 (1975), Nr. 11, S. 613–620. – ISSN 0001-0782

[130] Salton, Gerald: *Automatic Information Organization and Retrieval*. McGraw Hill, 1968

[131] Salton, Gerald ; Harrison, Michael A. (Hrsg.): *Automatic Text Processing*. Addison-Wesley, 1989

[132] Sarkar, I. N. ; Cantor, M. N. ; Gelman, R. ; Hartel, F. ; Lussier, Y. A.: Lin-

king biomedical language information and knowledge resources: GO and UMLS. In: *Pac Symp Biocomput* (2003), S. 439–50

[133] Schatz, B. R.: Information retrieval in digital libraries: bringing search to the net. In: *Science* 275 (1997), Nr. 5298, S. 327–34

[134] Schmid, Helmut: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *International Conference on New Methods in Language Processing*. Manchester, UK, 1994

[135] Schomburg, I. ; Chang, A. ; Ebeling, C. ; Gremse, M. ; Heldt, C. ; Huhn, G. ; Schomburg, D.: BRENDA, the enzyme database: updates and major new developments. In: *Nucleic Acids Res* 32 Database issue (2004), S. D431–3

[136] Schulze-Kremer, S.: Ontologies for molecular biology and bioinformatics. In: *In Silico Biol* 2 (2002), Nr. 3, S. 179–93

[137] Searls, D.B.: The language of genes. In: *Nature* 420 (2002), Nr. 6912, S. 211–7

[138] Shah, P. K. ; Perez-Iratxeta, C. ; Bork, P. ; Andrade, M. A.: Information extraction from full text scientific articles: where are the keywords? In: *BMC Bioinformatics* 4 (2003), Nr. 1, S. 20

[139] Silberman, Steve: Talking to strangers. In: *Wired* 8 (2000), Nr. 5

[140] Simmonds, H. A. ; Reiter, S. ; Nishino, T.: *The Metabolic and Molecular Bases of Inherited Disease*. Bd. 2: *Hereditary xanthinuria*. New York : McGraw-Hill, 1995

[141] Semantic Knowledge Representation Group / Lister Hill National Center for Biomedical Communications. URL http://mmtx.nlm.nih.gov/index.shtml, 2002. – Forschungsbericht

[142] Spector, K. S.: Diabetic cardiomyopathy. In: *Clin Cardiol* 21 (1998), Nr. 12, S. 885–7

[143] Stapley, B.J. ; Benoit, G.: Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In: *Pac Symp Biocomput* (2000), S. 529–40

[144] Stapley, B.J. ; Kelley, L.A. ; Sternberg, M.J.: Predicting the sub-cellular location of proteins from text using support vector machines. In: *Pac Symp Biocomput* (2002), S. 374–85

[145] Stephens, M. ; Palakal, M. ; Mukhopadhyay, S. ; Raje, R. ; Mostafa, J.: Detecting gene relations from Medline abstracts. In: *Pac Symp Biocomput* (2001), S. 483–95

[146] Steyvers, Mark ; Tenenbaum, Joshua B.: *The large-scale structure of semantic networks: Statistical analysis and a model of semantic growth*. – Submitted to Cognitive Science

[147] Strogatz, S. H.: Exploring complex networks. In: *Nature* 410 (2001), Nr. 6825,

S. 268–76

[148] Swanson, D. R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. In: *Perspect Biol Med* 30 (1986), Nr. 1, S. 7–18

[149] Swanson, D.R: Migraine and magnesium: Eleven neglected connections. In: *Perspectives in Biology and Medicine* 31 (1988)

[150] Tao, Yong-Chuan ; Leibel, Rudolph: Identifying functional relationships among human genes by systematic analysis of biological literature. In: *BMC Bioinformatics* 3 (2002), Nr. 1, S. 16. – URL http://www.biomedcentral.com/1471-2105/3/16. – ISSN 1471-2105

[151] Temkin, J. M. ; Gilder, M. R.: Extraction of protein interaction information from unstructured text using a context-free grammar. In: *Bioinformatics* 19 (2003), Nr. 16, S. 2046–53

[152] Thomas, J. ; Milward, D. ; Ouzounis, C. ; Pulman, S. ; Carroll, M.: Automatic extraction of protein interactions from scientific abstracts. In: *Pac Symp Biocomput* (2000), S. 541–52

[153] Tipton, K. ; Boyce, S.: History of the enzyme nomenclature system. In: *Bioinformatics* 16 (2000), Nr. 1, S. 34–40

[154] TouchGraph LLC / Open Source. URL http:/www.touchgraph.com, 2002. – Forschungsbericht

[155] Tsuruoka, Yoshimasa ; Tsujii, Jun'ichi: Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. In: Ananiadou, Sophia (Hrsg.) ; Tsujii, Jun'ichi (Hrsg.): *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, URL http://www.aclweb.org/anthology/W03-1306.pdf, 2003, S. 41–48

[156] Tyski, S. ; Colque-Navarro, P. ; Hryniewicz, W. ; Granstrom, M. ; Mollby, R.: Lipase versus teichoic acid and alpha-toxin as antigen in an enzyme immunoassay for serological diagnosis of Staphylococcus aureus infections. In: *Eur J Clin Microbiol Infect Dis* 10 (1991), Nr. 5, S. 447–9

[157] Valdés-Pérez, Raúl E.: Principles of human-computer collaboration for knowledge discovery in science. In: *Artificial Intelligence* 107 (1999), Nr. 2, S. 335–346. – ISSN 0004-3702

[158] Watts, D. J. ; Strogatz, S. H.: Collective dynamics of 'small-world' networks. In: *Nature* 393 (1998), Nr. 6684, S. 440–2

[159] Weeber, M. ; Klein, H. ; Aronson, A. R. ; Mork, J. G. ; Berg, L. T. de Jong-van den ; Vos, R.: Text-based discovery in biomedicine: the architecture of the DAD-system. In: *Proc AMIA Symp* (2000), S. 903–7

[160] Weeber, Marc: *Advances in literature-based discovery*. Online. 2003. – URL http://nl.ijs.si/et/talks/tsujiilab/saso/weeber.pdf

[161] Weiss, Sholom M. ; Apte, Chidanand ; Damerau, Fred J. ; Johnson, David E. ; Oles, Frank J. ; Goetz, Thilo ; Hampp, Thomas: Maximizing text-mining performance. In: *IEEE Intelligent Systems* 14 (1999), July/August, Nr. 4, S. 2–8

[162] Wilbur, W. J. ; Hazard, Jr. ; Divita, G. ; Mork, J. G. ; Aronson, A. R. ; Browne, A. C.: Analysis of biomedical text for chemical names: a comparison of three methods. In: *Proc AMIA Symp* (1999), S. 176–80

[163] Wilbur, W. J. ; Yang, Y.: An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. In: *Comput Biol Med* 26 (1996), Nr. 3, S. 209–22

[164] Wong, L.: PIES, a protein interaction extraction system. In: *Pac Symp Biocomput* (2001), S. 520–31

[165] Wren, J. D. ; Bekeredjian, R. ; Stewart, J. A. ; Shohet, R. V. ; Garner, H. R.: Knowledge discovery by automated identification and ranking of implicit relationships. In: *Bioinformatics* 20 (2004), Nr. 3, S. 389–98

[166] Wren, J. D. ; Garner, H. R.: Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. In: *Bioinformatics* 20 (2004), Nr. 2, S. 191–198

[167] Wuchty, S.: Scale-free behavior in protein domain networks. In: *Mol Biol Evol* 18 (2001), Nr. 9, S. 1694–702

[168] Yakushiji, A. ; Tateisi, Y. ; Miyao, Y. ; Tsujii, J.: Event extraction from biomedical papers using a full parser. In: *Pac Symp Biocomput* (2001), S. 408–19

[169] Yang, Yiming: An Evaluation of Statistical Approaches to Text Categorization. In: *Information Retrieval* 1 (1999), Nr. 1/2, S. 69–90

[170] Yeh, A. S. ; Hirschman, L. ; Morgan, A. A.: Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. In: *Bioinformatics* 19 Suppl 1 (2003), S. I331–I339

[171] Younkin, B. ; Gudzinowicz, B.: The viral mechanism of Reye's syndrome. In: *Med Hypotheses* 14 (1984), Nr. 2, S. 161–80

Glossar

Akronym	Eine Abkürzung, die sich aus den Anfangsbuchstaben mehrerer Wörter zusammensetzt.
Bewertungszahl	Durch *MetaMap* ermittelte Konfidenz in eine Konzeptzuordnung.
Disambiguierung	Auflösung von Mehrdeutigkeiten.
Enzymgraph	Netzwerk aus Enzymklassen, bei denen die Enzyme durch gemeinsame Krankheitskonzepte verbunden sind.
Flexion	Die Änderung oder Beugung eines Wortes mit seiner grammatikalischen Funktion.
Homonym	Ein Wort, dass für unterschiedliche Konzepte steht.
Knotengrad	Die Anzahl der Kanten eines Knoten in einem Netzwerk.
Kollokation	Die gemeinsame Nennung zweier Wörter in einem Textabschnitt.
Konzept	Eine Menge generischer, umfassender und akzeptierter Ideen. In dieser Arbeit ist mit einem Konzept ein mentales Bild, also ein Begriff gemeint.
Krankheitsgraph	Netzwerk aus Krankheitskonzepten, bei denen die Krankheiten durch gemeinsame Enzymklassen verbunden sind.
Morphem	Das kleinste bedeutungstragende Element einer Sprache.
n-Gram	Abfolge von n Worten.
Ontologie	Ein formales System von Konzepten, deren Bezeichnungen sowie semantischen Relationen zwischen den Konzepten.
Orthographie	Rechtschreibung.
Präzision	Der Anteil richtiger Ergebnisse im Verhältnis zu allen erhaltenen Ergebnissen.
Scale-Free-Graph	Netzwerk mit einer von der Größe des Netzwerks unabhängigen Verteilung der Knotengrade.

Small World-Netzwerk	Ein Netzwerk mit einem im Verhältnis zu seinem Gruppierungskoeffizienten kleinen mittlerem Knotenabstand.
Stammform	Die sinntragende Silbe eines Wortes, die übrig bleibt, wenn Vor- und Nachsilben entfernt werden.
Stoppwort	Ein Wort mit niedrigem Informationsgehalt. Stoppwörter werden vor dem Vergleich von Texten üblicherweise entfernt.
Synonym	Ein Synonym ist ein Wort, das in einem bestimmten Kontext die gleiche Bedeutung hat wie ein anderes Wort.
Textkorpus	Eine Sammlung von Dokumenten.
Vollständigkeit	Der Anteil richtiger Ergebnisse im Verhältnis zu allen richtigen Ergebnissen.
Zufallsgraph	Netzwerk mit einer zufälligen Verteilung der Knotengrade.

Curriculum Vitae

Adresse	Oliver Hofmann Klingelpütz 28 D-50670 Köln Telefon: +49-221-29 40 578 E-Mail: o.hofmann@smail.uni-koeln.de
Personalia	Geburtsdatum: 12. August 1972 Geburtsort: Köln Nationalität: Deutsch Familienstand: ledig

---

Schulbildung

8/1979	Grundschule Bodelschwingh, Hürth
8/1983	Albert-Schweitzer-Gymnasium, Hürth
6/1992	Abitur

Wehrdienst

10/1992–10/1993	Wehrdienst in Rheinbach

Studium

10/1993	Beginn des Biologiestudiums an der Universität Köln
10/1996	Vordiplomsprüfung
2/1999	Diplomprüfung, Hauptfach Genetik, Nebenfächer Biochemie und Entwicklungsbiologie
2/1999–3/2000	Diplomarbeit bei Frau PD Dr. R. Nischt an der Universitätsklinik Köln: *Analyse des humanen Nidogenpromotors*
seit 4/2000	Beginn der Doktorarbeit bei Herrn Prof. Dr. D. Schomburg an der Universität Köln
6/2001–12/2001	Forschungspraktikum bei deCODE Genetics, Reykjavik, Island

Köln, 18. Oktober 2004

Diese Dissertation wurde mit LaTeX unter Verwendung der Zeichensätze Adobe Sabon für den Fließtext, Linotype Syntax für Legenden und Euler für mathematische Zeichen gesetzt. Alle Diagramme wurden mit Gnuplot erstellt, in das METAPOST-Format umgewandelt und eingebunden. Abbildungen von Programmausgaben wurden vor der Integration in das Adobe PDF-Format umgewandelt. Zur Erstellung von Zeichnungen diente das XY-pic-Paket.