Entwicklung einer Textminingmethode zur automatisierten Extraktion von kinetischen Informationen aus der Literatur

I n a u g u r a l - D i s s e r t a t i o n
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Stephanie Eva Heinen
aus Köln
2007

Berichterstatter/in: Prof. Dr. Schomburg
Prof. Dr. Marner
Tag der letzten mündlichen Prüfung: 30.11.2007

Danksagung

Zuerst danke ich Herrn Prof. Dr. Schomburg für ein solch spannendes und interessantes Thema. Zudem möchte ich mich für die gute Betreuung bedanken, da er mir ein ständig verfügbarer Ansprechpartner bei jeder Art von Problem war. Auch bedanke ich mich für das Vertrauen und den Freiraum, dass es mir ermöglichte weitestgehend selbstständig zu arbeiten.

Ich möchte mich auch herzlich bei Prof. Dr. Marner dafür bedanken, dass er die zweite Korrektur dieser Arbeit übernommen hat.

Außerdem möchte ich mich bei meiner Familie (Mama, Papa, Gianni und meinem_Bärchen) bedanken, die nie an mir gezweifelt haben. Sie haben mich immer unterstützt und gefördert und haben mir sehr viel Kraft gegeben. Und ein ganz besonders liebes und großes Danke geht an meinen Bernie!

Last, but not least: ein großes Danke geht auch an Dr. Klaus Bollig, da er mir in jeder Situation hilfreich zur Seite stand.

Abkürzungen und Konventionen

Km: Michaelis-Menten Konstante (Schreibweise in der PubMed)

K_M: Michaelis-Menten Konstante (Schreibweise nach dem Système International d'Unités; SI)

Ki: Dissoziationskonstante des Enzym-Inhibitor-Komplexes (Schreibweise in der PubMed)

K_i: Dissoziationskonstante des Enzym-Inhibitor-Komplexes (Schreibweise nach dem Système International d'Unités; SI)

kcat: Wechselzahl (Schreibweise in der PubMed)

k_{cat}: Wechselzahl (Schreibweise nach dem Système International d'Unités; SI)

PDB: Protein Data Bank (http://www.rcsb.org/pdb/)

ID: Identifikationsnummer

Abstract: Kurzer zusammenfassender Auszug eines wissenschaftlichen Artikels

Token: wird als Synonym für einen Begriffe oder eine Zahl verwendet

Recall: Maß für die Anzahl extrahierter Informationen bezogen auf die ihre Menge im Gesamtbestand

Precision: Maß für die Relevanz bzw. Richtigkeit extrahierten Informationen

In dieser Arbeit werden die zu zitierenden Textauszüge in Anführungszeichen angegeben. Durch „…" werden Auslassungen in den Zitaten gekennzeichnet. Zahlen in eckigen Klammern geben die Nummer im Literaturanhang (siehe Abschnitt 6.) an.

Kurzzusammenfassung

Die Menge an verfügbaren biologischen Informationen ist über die letzten Jahre stetig angestiegen. Es ist daher essentiell, dass das enthaltene Wissen leicht zugänglich gemacht wird, wie z.B. in Datenbänken. Zum Erstellen solcher Datenbänke können Methoden zur automatischen Extraktion dieser Informationen verwendet werden. Eine pragmatische Methode zur Extraktion kinetischer Informationen aus ca. 17 Millionen Abstracts der PubMed unter Verwendung von Lexika wurde entwickelt. Es wurden Informationen zu K_M, K_i, k_{cat}, k_{cat}/K_M, V_{max}, IC_{50}, $S_{0.5}$, K_d, K_a, $t_{1/2}$, pI, Enzymnamen, EC Nummern, Liganden, Organismen, Lokalisationen, pH-Wert und Temperatur extrahiert.

509.153 kinetische Informationen konnten extrahiert werden, mit einer Precision von 55% bis zu 96% und einem Recall von 51% bis zu 84%. Die erhaltenen Informationen sind in der Datenbank „KID the KInetic Database“ im Internet zugänglich.

Abstract

The amount of available biological information is steadily increasing. It is therefore essential for this knowledge to be easily accessible. However much information is contained in the written literature and not in easy available ways, e.g. through databases, so that methods for the extraction of knowledge have to be applied. Text-mining is one way for the automatic accumulation of knowledge out of written text. A pragmatic approach for the extraction of kinetic information from about 17 million PubMed abstracts was developed. Information about K_M, K_i, k_{cat}, k_{cat}/K_M, V_{max}, IC_{50}, $S_{0.5}$, K_d, K_a, $t_{1/2}$, pI, enzyme-name, EC number, ligand, organism, localisation, pH and temperature was extracted.

509,153 kinetic parameters were found. A manual verification of the entries yielded to a precision from 55% up to 96% and a recall from 51% up to 84%. The attained information is stored in the database „KID the KInetic Database“, which is available via the internet.

Inhaltsverzeichnis

1. Einleitung

1.1. Begriffserklärung von Textmining

„Textmining ist ein automatisierter Prozess, mit dem natürliche Sprachen in Textform analysiert werden“ [9]. Es wurde [12, 13] als das automatische Auffinden neuer, zuvor unbekannter Informationen aus unstrukturierten Texten definiert. Oftmals wird es als Überbegriff für die drei Kategorien Auffinden von Informationen in relevanten Dokumenten („information retrieval“), Extraktion bestimmter Informationen („information extraction“) und „data mining“ aufgefasst [12, 13]. Unter letzterem versteht man das Aufdecken neuer Assoziationen zwischen den extrahierten Informationen [12, 13]. Auch Wissen, dass nur durch die Untersuchung mehrerer Textquellen erkennbar wird, soll ersichtlich werden [10]. Hierbei finden sowohl statistische als auch linguistische Verfahren Verwendung [11].

Aufgrund der erhaltenen Informationen können neue Arbeitsansätze entwickelt und Hypothesen verbessert werden [10]. Die erhaltenen Informationen können in vielen Fällen nicht nur extrahiert, sondern auch analysiert und interpretiert werden.

Ein gängiges Verfahren zur Identifizierung von gesuchten Begriffen ist die Verwendung einer lexikalischen Annäherung („dictionary approach“) [10, 12]. Hierbei werden zu suchende Begriffe in einem Lexikon gesammelt um sie später im Text wieder zu finden.

1.2. Bedeutung des Textminings

Die Menge an wissenschaftlichen Artikeln verdoppelt sich annähernd alle zehn Jahre [11]. Daher wird es für Wissenschaftler zunehmend schwierig selbst in eng umrissenen Arbeitsgebieten alle relevanten Veröffentlichungen zu verfolgen und zu berücksichtigen [14], so dass die Suche nach aktueller oder relevanter Literatur einen Großteil der Arbeitszeit des Forschers beanspruchen kann [11]. Besonders die Bereitstellung von Fachliteratur in

Onlinebibliotheken und deren verstärkte Nutzung (siehe Abbildung 1), wie z.B. der PubMed mit alleine 17 Mio. verfügbaren Artikeln, macht eine zeitsparende automatisierte Extraktion von Informationen wünschenswert [15]. In Anbetracht dessen gewinnt Textmining zunehmend an Bedeutung für Biologie, Chemie, Biochemie und Medizin [12, 14].

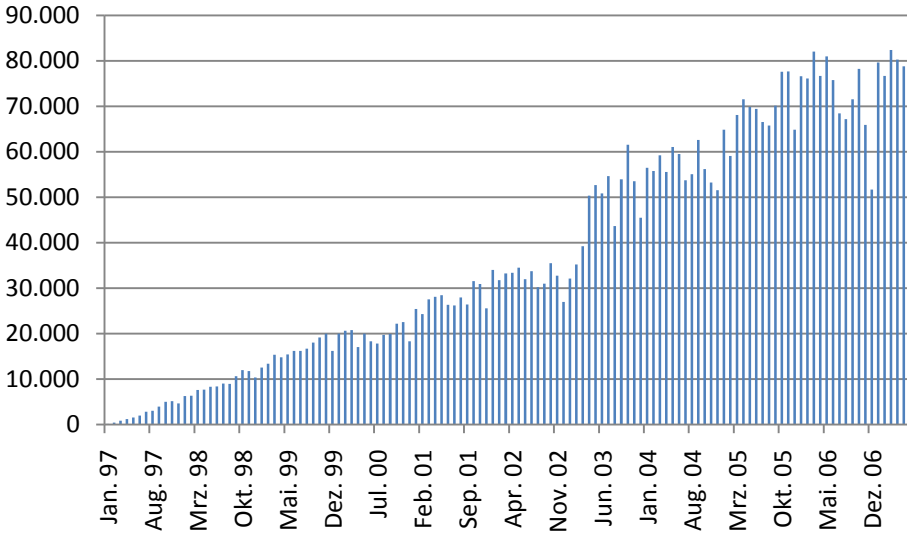

Abbildung 1: Statistik zur Nutzung der PubMed [22]

1.3. Definition und Bedeutung der Sprache

Sprache ist ein sich ständig weiterentwickelndes Produkt des gesellschaftlichen Lebens. Sie kann als System aus verbalen oder schriftlichen Zeichen verstanden werden, deren Ziel die zwischenmenschliche Kommunikation sowie der Austausch von Gefühlen, Gedanken und Wissen ist [19].

Die Regeln für den Gebrauch einer Sprache sind in der jeweiligen Grammatik definiert, können jedoch in den verschiedensten Formen ausgelebt werden (siehe den nachfolgenden Abschnitt). Zur Benennung der besprochenen Gegenstände dient die Verwendung von Entitäten [18], welche definierte deskriptive Zeichen oder Zeichenfolgen darstellen. So stellt die Zeichenfolge „alcohole dehydrogenase“ die sprachliche Kennung für das entsprechende Enzym dar.

1.4. Linguistische Anforderungen an Textminingprogramme

Lesen bezeichnet nicht nur das Erkennen von Buchstaben, sondern auch das Erfassen des Inhalts des geschriebenen Textes. Bereits Aristoteles entwickelte eine Theorie zu diesem Thema. Seine Hermeneutik (ca. 300 vor Christi) wird als die „Kunst der Interpretation von Aussagen“ [18] verstanden.

Das inhaltliche Erfassen von Texten stellt die größere Herausforderung bei der Entwicklung eines Textverarbeitungprogrammes dar. Hierzu müssen nicht nur Wörter und deren Bedeutung erkannt werden (Semantik) [11], sondern auch grammatikalische Regeln (Syntax) sind für die Perzeption des Inhaltes notwendig. Zusammenfassen lassen sich die diese Teilgebiete mit nachfolgend beschriebenen Pragmatik unter dem Begriff Semiotik [19].

Diese setzt voraus, dass die Grammatik nicht nur bekannt ist, sondern auch im Text richtig erkannt wird und in einen korrekten Kontext zur Wortebene gebracht wird. Es wird zusätzlich durch die Tatsache erschwert, dass Sprache auch eine menschliche Ausdrucksform darstellt. Je nach Autor lassen sich unterschiedliche Schreib- und Ausdrucksstile unterscheiden.

Dies ist auch bekannt als Freges Prinzip [11] und lässt sich wie folgt wiedergeben:

> „Die Bedeutung eines komplexen Ausdrucks ist eine Funktion der Bedeutung der Teile und der Art ihrer Zusammensetzung“ [11]

Wie sich die Bedeutung der einzelnen Teile auf die Gesamtaussage auswirkt zeigt sich im Vergleich der beiden folgenden syntaktisch analogen Sätze [11]. Da die beiden Sätze sich in ihrem letzten Teil unterscheiden, unterscheidet sich auch ihre Aussage.

- Der Hund beißt den Mann. [11]
- Der Hund beißt den Knochen. [11]

Den Einfluss der syntaktischen Zusammensetzung (Komposition) auf die Aussage des gesamten Satzes wird in dem folgenden Beispiel verdeutlicht [11]:

- Der Hund beißt den Mann. [11]
- Der Mann beißt den Hund. [11]

Hinzu kommt, dass der Inhalt eines Satzes durch verschiedene Formulierungen wiedergegeben werden kann (Synonymie) [28]. Je nach Formulierung kann sich die Position eines Wortes und sein grammatikalischer Fall vollständig verändern.

1. Paul hat ein schönes gelbes Fahrrad.
2. Das gelbe Fahrrad von Paul ist schön.
3. Paul hat ein gelbes Fahrrad und dieses gelbe Fahrrad ist schön.
4. Paul hat ein schönes, gelbes Fahrrad.
5. Paul hat ein Fahrrad, welches nicht nur gelb, sondern auch schön ist.
6. …

Dies wird zusätzlich erschwert durch die Mehrdeutigkeit einiger Wörter und Phrasen (Polysemie) [11, 12]. In solchen Fällen lässt sich die Bedeutung nur aus dem inhaltlichen Kontext der einzelnen Wörter und Phrasen im Satz erkennen.

1. Dieser Schimmel ist ein besonders schönes Tier.
2. Der Schimmel in der Wohnung hat zu schlimmen Erkrankungen geführt.

Die Herausforderung an den Programmierer besteht darin, jede der möglichen Formulierungen zu bedenken, interpretieren und für den Computer umsetzbar zu machen. Hier zeigt sich die Bedeutung der Pragmatik [11], welche die grammatikalisch analysierten Ausdrücke in Bezug auf den Interpretationskontext berücksichtigt. Sie bildet somit die „Schnittstelle“ zwischen inhaltlicher und grammatikalischer Analyse.

Hinzukommt, dass einige Informationen implizit in der Sprache enthalten sein können [18]. So enthält der Satz „In diesem Raum ist ein Nest aggressiver Wespen“ die implizite Information „Geh besser nicht in diesen Raum!“. Solche Informationen sind jedoch eher in der gesprochenen Form enthalten und können nach dem Stand der Technik nur schwer mit Hilfe der Computerlinguistik extrahiert werden.

1.5. Vorstellung der BRENDA

Die BRaunschweiger ENzyme DAtabase ist eine Onlinedatenbank für Enzyminformationen [26] und für den akademischen Gebrauch frei zugänglich [6, 7, 8]. Sie stellt Informationen aus verschiedensten Gebieten der Biologie zur Verfügung (siehe Tabelle 1). Der Großteil dieser Informationen wurde manuell aus der Literatur extrahiert.

Die Datenbank wurde 1987 in Braunschweig an dem National Research Center for Biotechnology gegründet [6, 7, 8]. Bis 2007 wurde sie an der Universität zu Köln geführt. Zurzeit wird sie wieder in Braunschweig verwaltet.

Tabelle 1: Informationsfelder der BRENDA und die Anzahl der enthaltenen Informationen [6]

Informationsfeld	Anzahl der Einzelinformationen
Enzymnamen und Synonyme	70.972
Reaktion und deren Spezifität	396.760
Funktionelle und kinetische Parameter	191.134
Literaturreferenzen	91.403

1.6. Bedeutung funktioneller Werte für Industrie und Forschung

Enzyme sind Biokatalysatoren mit einem pH- und einem Temperaturoptimum, die an fast allen biologischen Reaktionen teilnehmen und anhand ihrer kinetischen Eigenschaften charakterisiert werden können. Ihre katalytische Reaktion ist substrat- und produktspezifisch.

Ihre Bedeutung zeigt sich im Fehlen oder in einer Fehlfunktion im Organismus, was beides oftmals zu Krankheiten führt, so ist z.B. die Phenylketunorie (PKU) zurückzuführen auf einen genetisch bedingten Defekt der Phenylalanin Hydroxylase [2]. Ein Ansatz zur Heilung ist der Erwerb von Wissen über die Funktions- und Wirkungsweise des jeweils betroffenen Enzyms.

In manchen Fällen kann z.B. die Inhibition von Enzymen einen Ansatzpunkt zur Therapie darstellen. Aspirin (= Acetylsalicylsäure) beispielsweise hemmt Cyclooxygenasen, die die Produktion von Prostaglandinen steuern, und wirkt somit schmerzlindernd [27].

Ein oftmals zur Reinigung verwendetes Verfahren ist die Ionenchromatographie, bei der sich Proteine entsprechend ihres pI-Wertes aufreinigen.

Die Lagerungsdauer von Enzymen kann durch ihre Halbwertszeit beschrieben werden. Diese kann besonders für die Entwicklung von Produkten, deren Wirkkraft auf Enzymen beruht, z.B. bei Waschmittel, wichtig sein.

Auch bei Nachweisverfahren finden Enzyme oftmals Anwendung. So basieren z.B. Schwangerschaftstests auf ELISA (enzyme linked immune sorbent assay) [2].

1.7. Spezifizierung funktioneller Werte

Einen großen Teil der funktionellen Werte findet man im Gebiet der Enzymkinetik wieder. Diese beschreibt den zeitlichen Verlauf von Enzymreaktionen [1]. Sie kann unter anderem zur Aufklärung von Katalysemechanismen und deren Regulation dienen [1].

1.7.1. Dissoziations- und Assoziationskonstante

Die Dissoziationskonstante K_d ist durch folgende Gleichung gekennzeichnet:

$$E + A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} EA$$

$$K_d = \frac{k_{-1}}{k_1} = \frac{[A][E]}{[EA]}$$

A : Substrat
E : Enzym
EA : Enzymsubstratkomplex
k_1, k_{-1} : Geschwindigkeitskonstanten
Konzentrationen sind in eckigen Klammern [] angegeben.

Für spezifische Bindungen ist sie meist kleiner als 10^{-3} M [1]. Der Kehrwert wird als Assoziationskonstante K_a bezeichnet.

Für den Fall, dass einer der Reaktionspartner in einem solchen Überschluss vorliegt, dass seine Konzentration durch die Reaktion nicht merklich verändert wird, kann dies in die Dissoziationskonstante mit einbezogen werden. Dies trifft beispielsweise bei Reaktionen zu, bei denen Wasser einer der Reaktionspartner ist, z.B. bei der Hydrolyse.

$$A + H_2O \rightleftharpoons P + Q$$

P, Q : Hydrolyseprodukte

Ein solcher Prozess verändert die Wasserkonzentration nicht messbar, so dass die Reaktion behandelt werden kann, als ob Wasser nicht daran teilnehmen würde.

$$K'_d = \frac{[A][H_2O]}{[P][Q]} = K_d[H_2O]$$

$$K_d = \frac{[A]}{[P][Q]}$$

Ähnliches gilt für Protonen, die an einer enzymatischen Reaktion beteiligt sind. In diesem Fall wird die apparente Gleichgewichtskonstante K_{app} berechnet.

$$K_{app} = K_d[H^+]$$

Im Fall von Bindungsstudien an Ionenkanälen beschreibt die Dissoziationskonstante die Ionenkonzentration, bei der der Strom von Ladungen der Hälfte seines Maximums entspricht [20]. Die Strom- und Ionenkonzentration-Beziehung wird bei den meisten Kanälen durch eine Eins-zu-eins-Bindung erklärt, was bedeutet, dass jedes Ion für den Durchtritt einzeln an den Kanal binden muss. Trägt man in solchen Fällen den Strom gegen die Ionenkonzentration auf, ergeben sich verglichen mit der Enzym-Substrat-Bindung hohe K_d-Werte um 100 mM. Diese Werte deuten auf eine schwache Interaktion zwischen den beiden Reaktionspartnern hin, die typischerweise weniger als eine Mikrosekunde dauert [20]. Der hohe K_d-Wert ist somit eine Erklärung für eine hohe Leitfähigkeit von Kanälen.

Dissoziationskonstanten können auch auf Rezeptoren, wie z.B. Hormonrezeptoren, bezogen werden. In solchen Fällen wird K_d analog der Enzymreaktion definiert [21].

$$[H] + [R] \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} [HR]$$

[H] : Konzentration des Hormons
[R] : Konzentration des Rezeptors
[HR] : Konzentration des Hormon-Rezeptor-Komplexes

Hiermit ergibt sich:

$$K_d = \frac{k_{-1}}{k_{+1}}$$

Nach einer Umformung folgt:

$$\frac{[HR]}{[Rg]} = \frac{[H]}{(K_d + [H])}$$

Rg : Gesamtkonzentration des Rezeptors

Diese Gleichung kann zu einer Geradengleichung umgeformt werden:

$$\frac{[HR]}{[H]} = \frac{-1}{K_d \cdot [HR]} + \frac{[Rg]}{K_d}$$

Trägt man [HR]/[H] gegen [HR] auf, erhält man einen Scatchard-Plot (siehe Abbildung 2). -1/K_d entspricht dabei der Steigung der Graden.

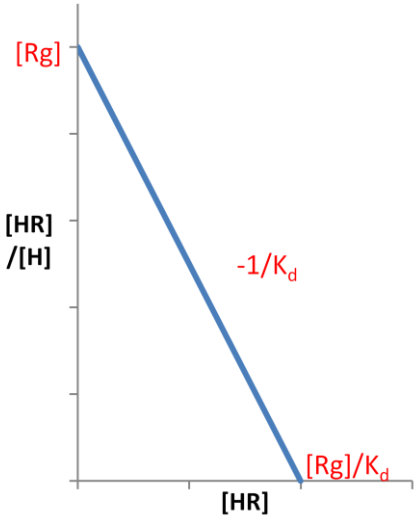

Abbildung 2: Scatchard-Plot für das Beispiel Hormon-Rezeptor-Bindung. -1/K_d entspricht der Steigung; [H] ist die Hormonkonzentration; [Rg] ist die Gesamtkonzentration des Rezeptors; der Abszissenschnittpunkt liegt bei [Rg]/K_d, der Ordinatenschnittpunkt liegt bei [Rg].

Die Dissoziationskonstante liegt im Fall von Hormonrezeptoren meist bei 10^{-11} und 10^{-8} mol/l [21].

In pharmakologischen Versuchen berechnet sich K_d ebenfalls aus dem Massenwirkungsgesetz, allerdings sind hier die Mengen an Pharmakon, freiem Rezeptor und Pharmakon-Rezeptor-Komplex nicht direkt messbar [43]. Die Gesamtrezeptorkonzentration (R_T) wird als konstant angenommen. Da die Menge an Pharmakonmolekülen in der Regel im Überschuss vorliegt, kann die eingesetzte Konzentration des Pharmakons [P] an Stelle der nicht gebundenen Pharmakonkonzentration [P_f]verwendet werden [43].

$$K_d = \frac{[P] \cdot ([P_T] - [PR])}{[PR]}$$

[P] : Konzentration des Pharmakons
[R] : Gesamtkonzentration des Rezeptors
[PR] : Konzentration des Pharmakon-Rezeptor-Komplexes

Durch einige Umformungen lässt sich die Konzentration von besetzten Rezeptoren in Abhängigkeit zur Pharmakonkonzentration beschreiben.

$$[PR] \cdot K_d = [P] \cdot ([R_T] - [PR])$$

$$[PR] \cdot K_d + [PR] \cdot [P] = [P] \cdot [R_T]$$

$$[PR] = \frac{[P] \cdot [R_T]}{K_d + [P]}$$

Somit ergibt sich bei [P] << K_d:

$$PR \cong [P] \cdot \frac{[R_T]}{K_d}$$

Bei geringen Pharmakonkonzentrationen nimmt [PR] linear mit [P] zu.

Wenn [P] sehr viel größer ist als K_d ergeben sich [PR] als ungefähr gleich groß im Vergleich mit [R_T].

$$[PR] \cong [R_T]$$

1.7.2. Michaelis-Menten-Kinetik

Bei der einfachsten anzunehmenden Enzymreaktion wird nur ein Substrat umgesetzt. Es wird angenommen, dass die Reaktion irreversibel sei. Die Gleichung für diese Reaktion lautet:

$$A + E \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} EA \overset{k_2}{\rightarrow} E + P$$

A : Substrat
E : Enzym
P : Produkt
EA : Enzymsubstratkomplex
k_1, k_{-1},k_2: Geschwindigkeitskonstanten

Betrachtet man die zeitliche Veränderung der einzelnen Reaktionspartner, dann ergeben sich folgende Differentialgleichungen:

$$\frac{d[A]}{dt} = -k_1[A][E] + k_{-1}[EA] \qquad 1.1$$

$$\frac{d[E]}{dt} = -k_1[A][E] + (k_{-1} + k_2)[EA] \qquad 1.2$$

$$\frac{d[EA]}{dt} = k_1[A][E] - (k_{-1} + k_2)[EA] \qquad 1.3$$

$$\frac{d[P]}{dt} = k_2[EA] = \upsilon \qquad 1.4$$

Die Produktzunahme wird in der letzten Formel mit der Umsatzgeschwindigkeit υ gleichgesetzt. Sie ist der Konzentration des Enzym-Substrat-Komplexes EA direkt proportional.

Für den Fall eines Fließgleichgewichtes (steady-state) gleichen sich Bildung und Zerfall des Enzym-Substrat-Komplexes aus. Die Substratabnahme und Produktzunahme sind in dieser Phase linear. Die zeitliche Änderung des Enzym-Substrat-Komplexes und des freien Enzyms kann null gesetzt werden, so dass eine Reaktion nullter Ordnung vorliegt.

$$\frac{d[EA]}{dt} = \frac{d[E]}{dt} = 0$$

Die Gleichungen 1.2 und 1.3 vereinfachen sich somit zu folgendem Term:

$$k_1[A][E] = (k_{-1} + k_2)[EA]$$

Ersetzt man [E] durch $[E] = [E_0] - [EA]$, dann ergibt sich die Konzentration des Enzym-Substrat-Komplexes wie folgt:

$$[EA] = \frac{k_1[A][E]_0}{k_1[A] + k_{-1} + k_2}$$

Setzt man den obenstehenden Term in die Formel für die Umsatzgeschwindigkeit υ (1.4) ein, dann ergibt sich folgender Ausdruck:

$$\upsilon = \frac{d[P]}{dt} = k_2[EA] = \frac{k_2[E]_0[A]}{\frac{k_{-1} + k_2}{k_1} + [A]}$$

Die Michaelis-Menten-Konstante wird definiert als:

$$K_M = \frac{(k_{-1} + k_2)}{k_1}$$

Unter Verwendung von K_M und der Maximalgeschwindigkeit $V_{max} = k_2[E]_0$ lässt sich die unten stehende Formel erstellen.

$$v = \frac{V_{max}\,[A]}{K_M + [A]}$$

Diese Gleichung ist unter der Bezeichnung Michaelis-Menten-Gleichung bekannt.

Für den Fall, dass k_2 sehr viel kleiner ist als k_{-1}, ergibt sich $K_M = \frac{k_{-1}}{k_1}$. K_M stellt die Dissoziationskonstante des Enzym-Substrat-Komplexes dar und ist dann ein Maß für die Affinität zwischen Enzym und Substrat. Ein kleiner Wert weist auf eine hohe Affinität hin.

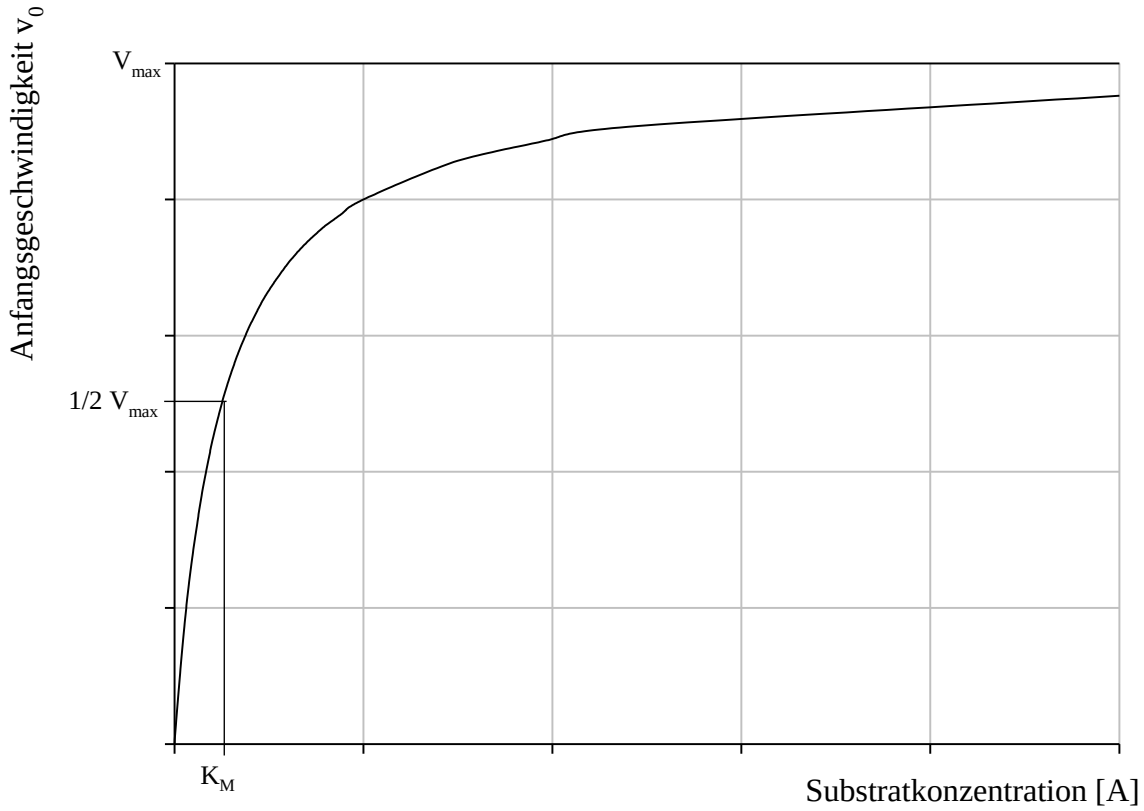

Abbildung 3: Graphische Darstellung einer Michaelis-Menten-Kinetik [2]. V_{max}: Maximalgeschwindigkeit; K_M: Michaelis-Menten Konstante.

Anhand von Abbildung 3 lässt sich K_M auch graphisch ermitteln. Bei K_M handelt es sich um die Substratkonzentration bei halbmaximaler Geschwindigkeit. Gewöhnlich liegen die Werte zwischen 10^{-1} und 10^{-7} M [2].

V_{max} gibt die maximale Geschwindigkeit der enzymatischen Reaktion an. Aus dem folgenden Term ergibt sich, dass die Geschwindigkeit abhängig von der Enzymkonzentration und dem geschwindigkeitsbestimmenden Schritt der Reaktion ist.

$$V_{max} = k_{cat} \cdot [E]_0 = k_2 \cdot [E]_0$$

Die Konstante k_{cat} [s^{-1}] ist ein Maß für die Umsatzgeschwindigkeit eines Enzyms. Sie wird auch „turnover number“ oder Wechselzahl genannt und gibt an, wie viele Substratmoleküle in einer gegebenen Zeiteinheit von einem einzigen Enzymmolekül in das entsprechende Produkt umgesetzt werden [2]. Aus der Gleichung für die Geschwindigkeit ergibt sich für k_{cat} folgende Formel:

$$k_{cat} = \frac{V_{max}}{[E]_0}$$

Der Ausdruck katalytische Effizienz beschreibt den folgenden Term und hat die Dimension einer Geschwindigkeitskonstanten zweiter Ordnung:

$$\frac{k_{cat}}{K_M} = \frac{k_{cat}\, k_1}{(k_{-1}+k_2)}$$

Ein hoher Wert weist auf eine hohe Substratspezifität hin. Durch die Diffusion von Molekülen im Raum wird dieser Werte auf 10^8 und 10^9 mol^{-1} s^{-1} [2] begrenzt.

1.7.3. Spezifische Aktivität

Die spezifische Aktivität gibt die Menge von umgesetztem Substrat pro Milligramm Enzym an. Die Einheit Unit ist definiert als die Enzymmenge, die ein µMol Substrat pro Minute unter Standardbedingungen umsetzt. Die SI-Einheit heißt katal (Symbol: kat) und ist definiert als die Enzymmenge, die ein Mol Substrat in einer Sekunde umsetzt (1 kat = 6 x10^7 Units). Die spezifische Aktivität ist jeweils bezogen auf die eingesetzte Enzymmenge. Sie wird gewöhnlich in Units/mg angegeben.

1.7.4. Inhibitionskinetik

Inhibition stellt den negativen Einfluss bestimmter Liganden auf die Aktivität des jeweiligen Enzyms dar. Diese Stoffe heißen Inhibitoren und können je nach Art der Hemmung im aktiven Zentrum oder an eigenen Bindungsstellen binden. Es kann zwischen reversibler und irreversibler Inhibition unterschieden werden.

Die Reaktionsgleichung für die Inhibition lässt sich in folgender Formel zusammenfassen:

$$E + I \rightleftharpoons EI$$

I : Inhibitor
E : Enzym
EI : Enzyminhibitorkomplex

Die Inhibitionskonstante K_i ist durch den untenstehenden Term definiert.

$$K_i = \frac{[E][I]}{[EI]}$$

Durch die kinetischen Eigenschaften lassen sich Rückschlüsse auf die Art der Inhibition ziehen. Dies wird in Abbildung 4 in einer doppeltreziproken Darstellung verdeutlicht.

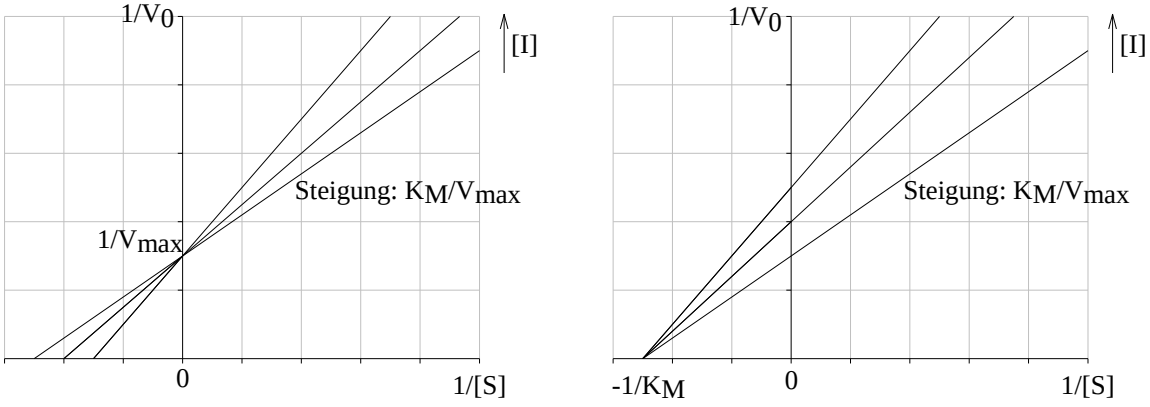

Abbildung 4: doppeltreziproke Darstellungen kompetitiver (links) und nichtkompetitiver Hemmung (rechts) [2]. V_{max}: Maximalgeschwindigkeit; K_M: Michaelis-Menten Konstante; [I]: Inhibitorkonzentration; [S]: Substratkkonzentration; V_o: Anfangsgeschwindigkeit.

Bei der kompetitiven Hemmung bleibt V_{max} gleich und K_M nimmt mit steigender Konzentration des Inhibitors zu. Bei der nicht-kompetitiven Hemmung wird V_{max} verringert, aber K_M bleibt unverändert.

Der IC_{50}-Wert gibt die Inhibitorkonzentration an, die nötig ist, um die Enzymaktivität in vitro um 50% zu verringern. Misst man diese Konzentration in vivo, spricht man von EC_{50}.

Im Fall einer kompetitiven Inhibition ist dieser Wert abhängig von K_M, K_i und der Substratkonzentration [41].

$$IC_{50} = \left[1 + \frac{[Substrat]}{K_M}\right] \cdot K_i$$

Beide Werte liegen meist im nanomolaren oder mikromolaren Bereich.

Oft findet auch der pIC_{50} Verwendung:

$$pIC_{50} = -\log IC_{50}$$

1.7.5. Kooperativität und Hill-Kinetik

Als Kooperativität bezeichnet man Wechselwirkungen verschiedener Untereinheiten in einem Enzym.

Die Hill Gleichung beschreibt die kooperative Wirkung bei der Bindung des Liganden. In solchen Fällen zeigen die Bindungskurven einen typischen sigmoidalen Verlauf. Erstmals entdeckt wurde Kooperativität bei der Bindung von Sauerstoff an Hämoglobin. Bei diesem aus vier Untereinheiten bestehendem Protein stimuliert die Bindung des ersten O_2-Moleküls die Bindung der weiteren Substratmoleküle. Bei Enzymen wird Kooperativität oft zur Regulation der Katalyse verwendet [1].

1910 postulierte A. V. Hill, dass mehrere (n) Sauerstoffmoleküle (A) gleichzeitig an Hämoglobin (E) binden können [1].

$$E + nA \leftrightharpoons EA_n$$

Hieraus ergibt sich:

$$\mathrm{K_d} = \frac{[E][A]^n}{[EA_n]}$$

Die Bindungsgleichung lautet in diesem Fall:

$$r = \frac{n[A]^n}{K_d + [A]^n}$$

Diese Formel ist unter dem Namen Hill-Gleichung bekannt und beschreibt sigmoide Bindungskurven (s. Abbildung 5).

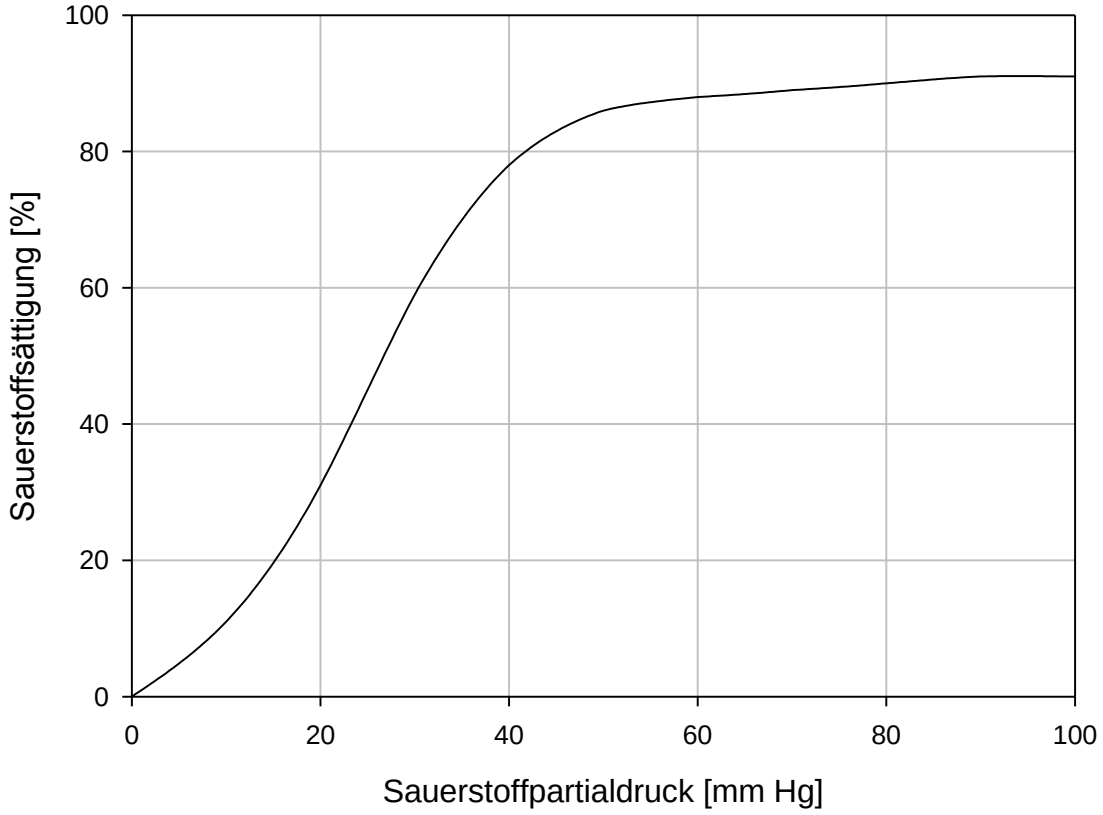

Abbildung 5: Sättigungskurve für Hämoglobin und Sauerstoff [1]

Durch Logarithmieren wird die oben genannte Gleichung in eine lineare Form umgewandelt, wobei r durch $\overline{\gamma} = r/n$ ersetzt wird.

$$\frac{\overline{\gamma}}{1-\overline{\gamma}} = \frac{[A]^n}{K_d}$$

$$\log\frac{\overline{\gamma}}{1-\overline{\gamma}} = n \cdot log[A] - logK_d$$

A : Substrat

$\overline{\gamma}$: r/n

n : Anzahl der Substratmoleküle

Durch das logarithmische Auftragen von $\overline{\gamma}$ / $(1-\overline{\gamma})$ gegen die Konzentration des Liganden lässt sich die Anzahl der Bindungsstellen an einem Protein mit nicht kooperativen Bindungszentren ermitteln. n ist mit der Steigung der Geraden gleichzusetzten. Eine solche Auftragung wird Hill-Diagramm genannt (siehe Abbildung 6).

Für Hämoglobin und andere Enzyme lassen sich die ermittelten Daten nicht in Form einer Geraden darstellen. Das Diagramm weist in diesen Fällen einen charakteristischen

dreiphasigen Verlauf auf. Die Kurve verläuft für geringe Substratkonzentrationen linear und zeigt eine Steigung von 1. Für Hämoglobin erhält man für den mittleren Sättigungsbereich einen Maximalwert von 2,8. Für den Sättigungsbereich lässt sich wiederum eine Gerade mit einer Steigung von 1 ermitteln.

Für hyperbolisch verlaufende Bindungskurven, wie beispielsweise für Myoglobin, beträgt die Steigung in solchen Auftragungen immer 1 (siehe letztgenannte Formel).

In dem Fall einer positiven Kooperativität ist der Hill-Koeffizient n_H grösser als 1, d.h. die Funktionen der Untereinheiten verstärken sich gegenseitig. Bei einer negativen Kooperativität ist der Koeffizient kleiner als 1.

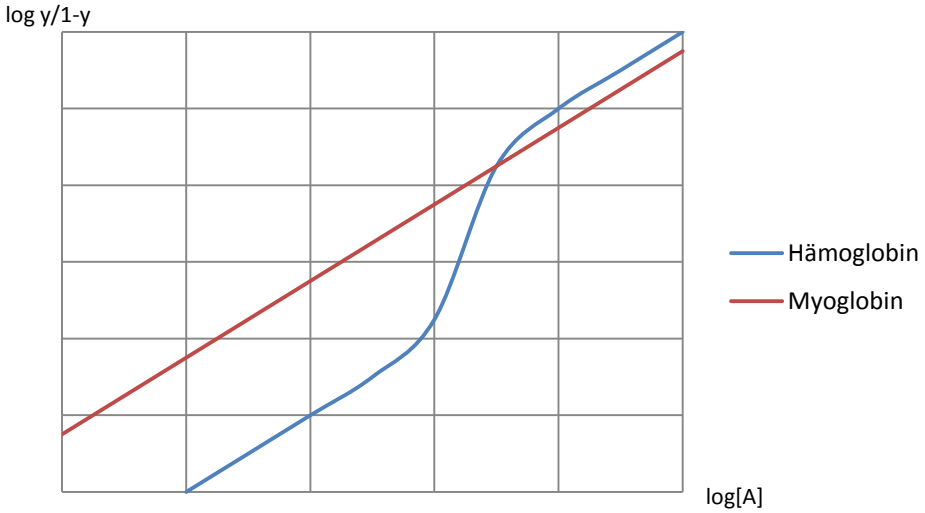

Abbildung 6: Hill-Diagramm für Hämoglobin mit dreiphasigem Verlauf im Vergleich mit Myoglobin [1].

1.7.6. Allosterische Enzyme

Effektoren, d.h. Aktivatoren und Inhibitoren, besitzen in manchen Enzymen eigene Bindungsstellen. Diese Bindungsstellen werden als allosterische Zentren bezeichnet. Durch die Bindung des Effektors an dieses Zentrum wird die Konformation des aktiven Zentrums beeinflusst und eine spezifische Regulation gewährleistet.

Die Bindungskurven dieser Enzyme weisen einen sigmoiden Verlauf auf und folgen somit nicht der Michaelis-Menten-Kinetik, deren Enzyme stets einen hyperbolischen Verlauf zeigen. Für ihre kinetische Beschreibung ist eine andere Terminologie entwickelt worden. So wird statt K_M die Bezeichnung $S_{0.5}$ oder $K_{0.5}$ verwendet.

1.7.7. Weitere Enzyminformationen

Der isoelektrische Punkt beschreibt den pH-Wert, an dem das Enzym keine elektrische Nettoladung aufweist [2]. Experimentell kann dieser Wert über eine Titrationskurve ermittelt werden. Oft wird dieser Wert auch als pI oder pH_I bezeichnet.

Der Begriff $t_{1/2}$ ist eine Abkürzung für verschiedene Halbwertszeiten, so wird er z.B. in der Atomphysik als Lebenszeit von Isotopen und in der Pharmakologie als Verweilzeit von Medikamenten im Organismus verwendet. In dieser Arbeit gibt die Halbwertszeit $t_{1/2}$ die Hälfte der Zeit an, nach der ein Enzym oder Protein seine Funktionsfähigkeit verliert (Beispiele siehe Tabelle 2).

Tabelle 2: Halbwertszeiten von Proteinen [45]

Enzym	Halbwertszeit
Ornithine decarboxylase	0,2 Stunden
Pyruvate kinase	30 Stunden
Aldolase	118 Stunden
Cytochrome c	150 Stunden

1.8. Zielsetzung

Ziel der Doktorarbeit ist die Entwicklung eines Textminingprogramms zur automatischen Extraktion von funktionellen, insbesondere kinetischen, Informationen aus der Literatur. Hierbei wurde nach K_M, K_i, k_{cat}, V_{max}, K_d, n_H, IC_{50}, $S_{0.5}$, der spezifischen Aktivität, der Halbwertszeit $t_{1/2}$ und dem isoelektrischen Punkt pI gesucht. Diese Parameter sollten mit Informationen über den Organismus, das Enzym, die EC Nummer, den Liganden, das Gewebe, den pH-Wert und die Temperatur komplettiert werden.

Die BRENDA stellt bereits viele dieser Werte zur Verfügung. Bislang erfolgt die Annotation neuer Informationen jedoch weitgehend manuell. Durch die Verwendung dieses Computerprogrammes soll dies für die obengenannten Werte beschleunigt werden.

Zudem soll eine Datenbank mit Informationen zu Enzymen erstellt werden, deren Zielgruppe analytisch tätige Forscher sind. Zusätzlich soll ein schneller Zugriff auf die originale Textquelle ermöglicht werden. Diese Datenbank eignet sich, anhand der gesammelten Informationen, zur Bestätigung, Verbesserung oder Weiterentwicklung von Experimenten.

2. Material und Methoden

Die Prozessierung der Abstracts umfasst mehrere Teilschritte, die in der folgenden Abbildung (Abbildung 7) dargestellt werden. Zu Beginn wird der Abstract in Sätze getrennt, in denen unter Verwendung von Lexika zu suchende Werte identifiziert werden. Die extrahierten Werte werden anschließend entsprechen ihrem inhaltlichen Zusammenhang im Abstract verbunden. Abschließend erfolgen die Normierung der ermittelten Werte und die Präsentation der Ergebnisse. Jeder Teilschritt wird in den folgenden Unterpunkten des Kapitels näher behandelt.

Abbildung 7: Fließschema des Programmablaufs.

2.1. Verwendete Programme und Techniken

Für den im Folgenden präsentierten Algorithmus wurden die Programmiersprachen C++, C#, Qt4 [24], Python und MySQL mit den Entwicklungsumgebungen Visual Studio und IDL verwendet. Die Entwicklung des Onlineinterfaces erfolgte unter Verwendung von Joomla [25], PHP und JavaSkript.

2.2. Sammeln der Daten

2.2.1. Herunterladen der PubMed

Unter Verwendung des SOAP-Interfaces [46] der PubMed werden die Abstracts mit den zugehörigen Informationen (Titel, Autor, Meshterms, Journal, Ausgabe, Seitenzahlen, Erscheinungsdatum) automatisch heruntergeladen und in einer lokalen Kopie gespeichert.

2.2.2. Erstellung der Lexika

Für die Identifizierung gesuchter Ausdrücke werden verschiedene Lexika mit Hilfe von Tabellen aus der BRENDA erstellt. Es werden Lexika für Enzymnamen, EC-Nummern, Liganden, Gewebe, Organismen erstellten. Um Polysemie (s. Abschnitt 1.4) bei den Ausdrücken zu verhindern, werden manuell erstellte Ausschlusslisten verwendet. So kann z.B. die Abkürzung IMP sowohl für einen Liganden (Inositol monophosphate) wie auch für ein Enzym (Inositol-phosphate phosphatase; EC 3.1.3.25) stehen.

Zusätzlich werden manuell erstellte Listen mit Einträgen, die nicht in der BRENDA enthalten sind (z.B. implizites Gewebe bei „renal“), in die Lexika integriert.

Für Gewebe wird zusätzlich eine Tabelle der UniProt [23] verwendet.

Die Lexika für die verschiedenen kinetischen Ausdrücke (K_M, K_i, k_{cat}, V_{max}, $S_{0.5}$, IC_{50}, k_{cat}/K_M, K_d, K_a, pI, $t_{1/2}$, n_H, spezifische Aktivität, pH, Temperatur) und ihre Einheiten werden manuell mit Hilfe der Abstracts erstellt.

2.3. Wertextraktion

2.3.1. Satzerzeugung

Zur Unterscheidung einzelner Sätze im Abstract wurde eine Definition des Satzendes als Punkt gefolgt von einem Leerzeichen („. “) durchgeführt. Es wird weiterhin ausgeschlossen, dass es sich an der Trennungsstelle im Satz nicht um eine Abkürzung (z.B. „e.g. “) oder eine Namensinitiale (z.B. „A. Smith“) handelt. „C“ und „M“ werden von den Initialen ausgenommen, da sie auch für Temperaturen und Konzentrationen stehen können.

Desweiteren soll die Anzahl der sich öffnenden Klammern im Satz („(“ und „[“) der Anzahl der sich schließenden Klammern („)“ und „]“) entsprechen.

Abschließend erfolgt eine Überprüfung des Satzanfangs. Dieser darf nicht mit alleinstehenden Einheiten wie z.B. („s-1“ oder „cm-1“) beginnen. Auch konjugierte Formen einiger Verben sollen nicht am Satzanfang stehen („is“, „was“, „were“, „has“, „have“, „had“). Somit können alle Formen von konjugierten Verben im Futur, Perfekt und Plusquamperfekt ausgeschlossen werden („being“ und „having“ werden am Satzanfang erlaubt).

2.3.2. Tokenisierung und Identifizierung

Für die Tokenisierung [11] werden die Sätze zunächst an den Leerzeichen in Teiltoken getrennt. Die Anfangs- und Endposition bezogen auf die absolute Zeichenposition im Abstract, ein eventuell vorhandenes endständiges Komma (das bei der späteren inhaltlichen Verbindung (siehe Abschnitt 2.4) von Bedeutung ist) und die Zeichenfolge, werden in einem Objekt vermerkt.

Die Lexika werden aus der Datenbank eingelesen und die einzelnen Einträge in eine Abfolge von Teiltoken umgewandelt. Um eine schnelle Verarbeitung der Daten zu ermöglichen, wird das folgende Prinzip auf der Grundlage hashbasierter Lexika mit konstanter Laufzeit angewendet (vergleiche Abbildung 8). Ausgehend von dem ersten Teiltoken (z.B. „glucose“) folgt jeweils eine Auswahl verschiedener Teiltoken (z.B. „6-“, „phosphatase“ und „6-phosphate“), woran sich wiederum eine Auswahl anschließen kann (z.B. „phosphate“ und „acetate“). Das letzte Teiltoken einer Reihe von Teiltoken trägt eine Markierung mit der jeweiligen Suchkategorie (z.B. Enzym oder Ligand).

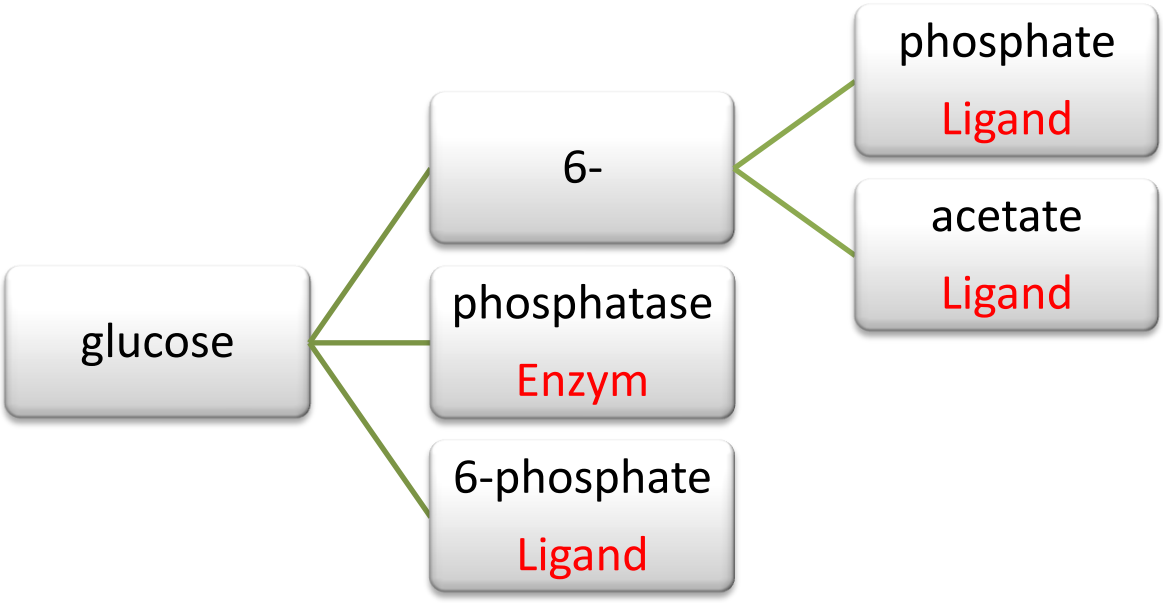

Abbildung 8: Schematische Darstellung des Lexikonaufbaus; Markierungen zur Unterscheidung einzelner Kategorien für die Suche sind in rot markiert

Für die Identifizierung werden die Token im Satz der Reihe nach durchgegangen. Ist ein Token in dem Lexikon enthalten, werden die nachfolgen Token auch auf ihre Anwesenheit im Lexikon geprüft (s. Abbildung 9). Die längste Abfolge von Einzeltoken mit einer Markierung wird im Satz zu einem gemeinsamen Token zusammengefasst und mit der jeweiligen Suchkategorie (z.B. Ligand) markiert.

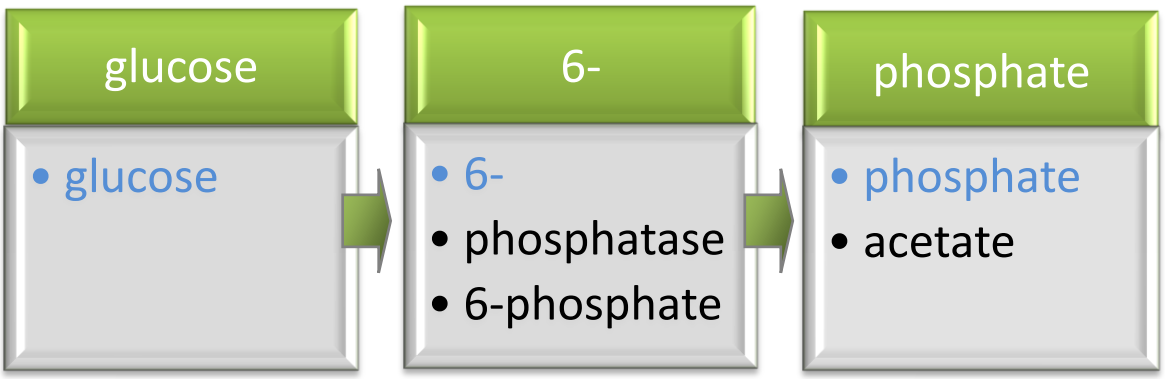

Abbildung 9: Schematischer Ablauf der Identifizierung; die grün-unterlegten Felder stellen die verschiedenen Einzeltoken eines Liganden im Satz dar. Die grau-unterlegten Felder stellen die jeweiligen Auswahlmöglichkeiten im Lexikon dar. Die blau-markierte Auswahlmöglichkeit stimmt mit der Abfolge der Einzeltoken im Satz überein.

In den Sätzen, die eine kinetische Variable enthalten, werden Zahlen über reguläre Ausdrücke identifiziert (s. Tabelle 6). Wird ein Ausdruck für eine Zahl von einer Einheit gefolgt, so werden Zahl und nachstehende Einheit zu einem gemeinsamen Token zusammengefasst.

Stehen bestimmte Markierungen vor oder hinter einer Zahl (z.B. eine nicht relevante Einheit), wird die Zahl im Satz gelöscht, was deren Gesamtzahl reduziert und das Risiko einer falschen Verbindung verringert werden.

Steht ein Ligand hinter einer negierenden Phrase (z.B. „..in absence of ATP…“) wird er aus den Einträgen für die inhaltliche Verbindung entfernt.

Weist ein Token ein endständiges Komma auf, wird aber von einer Markierung wie „which“ oder „whose“ gefolgt, so wird der Vermerk für das Komma im Token gelöscht. Bei der späteren inhaltlichen Verbindung kann somit über das Komma hinweg in die nächste Satzebene hinein gelesen werden (s. Abschnitt 2.4.1.1)

2.4. Inhaltliches Verbinden der identifizierten Entitäten

Für die inhaltliche Verbindung der identifizierten Entitäten wird sowohl eine direkte als auch indirekte Verbindung verwendet (siehe Abschnitt 2.4.1 und 2.4.2).

2.4.1. Direkte inhaltliche Verbindung

Die Position des kinetischen Ausdrucks bildet den Startpunkt für die direkte Verbindung. Von dieser Position aus wird zuerst nach rechts und anschließend nach links verbunden.

Zwei Entitäten unterschiedlicher Kategorien werden miteinander verbunden, wenn sie im Satz direkt nebeneinander stehen oder die Token zwischen ihnen als verbindende Ausdrücke gekennzeichnet worden sind (s. Abbildung 10). Junktoren („and“ und „or“), die nicht Teil einer Auflistung sind (s. Abschnitt 2.4.1.2), endständige Semikolons und endständige Kommata werden in diesem Fall als Terminationssignal interpretiert. Die direkte Verbindung wird ebenfalls beim Erreichen des Satzanfangs bzw. Satzendes beendet.

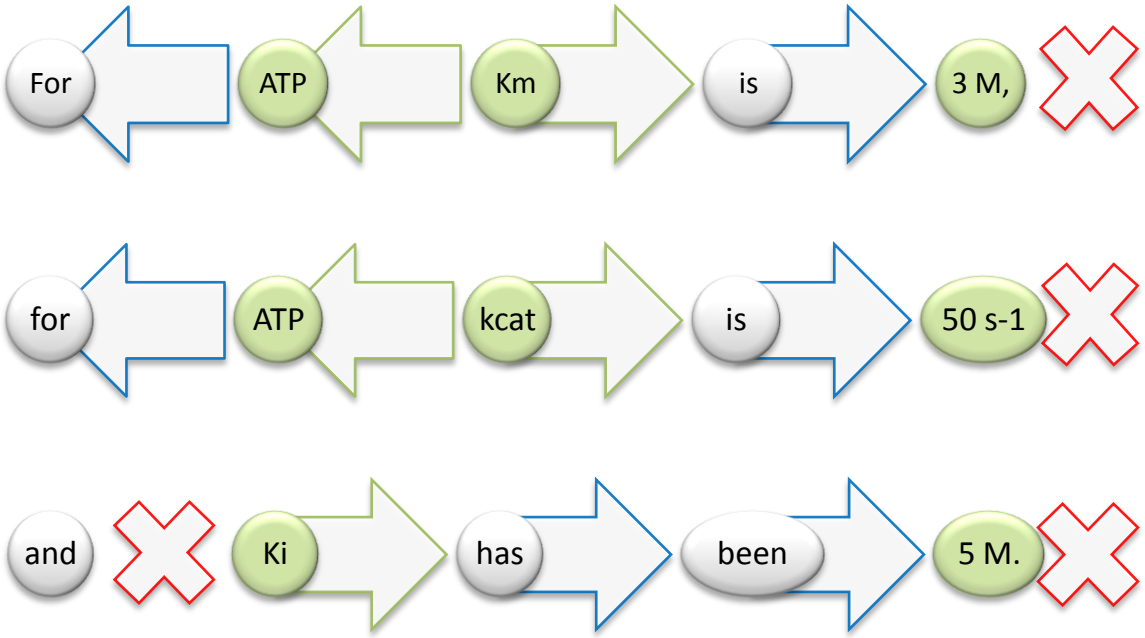

Abbildung 10: Schema der inhaltlichen direkten Verbindung anhand eines Beispielsatzes; identifizierte Entitäten sind mit einem grünen Kreis unterlegt; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren zusätzlich verbindende Ausdrücke; rote Kreuze beenden die Verbindung bei Kommata, Satzende oder fehlenden verbindenden Ausdrücken

2.4.1.1. Auflösen von Satzebenen

In Fällen, in denen ein Token mit einem endständigen Komma von einem definierten Schlüsselwort, wie „which“, gefolgt wird (s. Abschnitt 2.3.2), wird der folgende Nebensatz in die inhaltliche Verbindung einbezogen. In dem Beispiel in Abbildung 11 wird somit die Verbindung zu dem Liganden „alcohol“ korrekt ermöglicht.

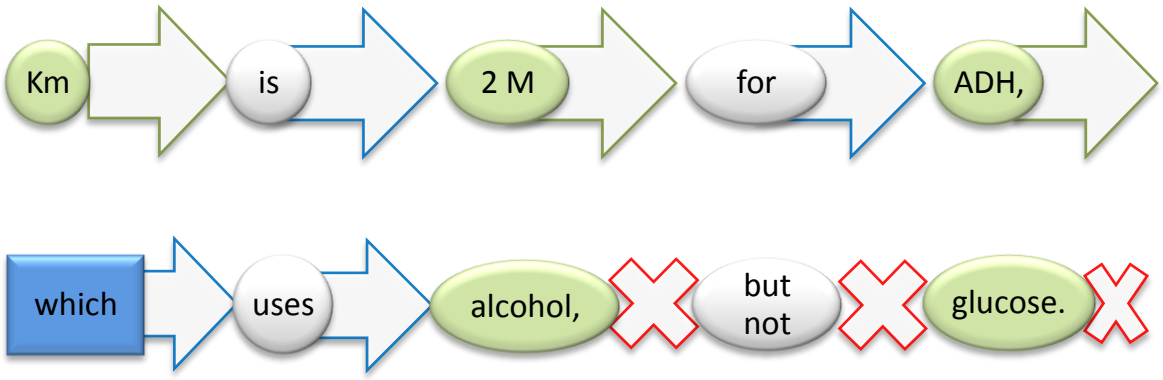

Abbildung 11: Schema der inhaltlichen direkten Verbindung anhand eines Beispielsatzes mit einbezogenem Nebensatz; identifizierte Entitäten sind mit einem grünen Kreis unterlegt; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren zusätzlich verbindende Ausdrücke; Schlüsselworte zum Ignorieren des Kommas sind blau unterlegt; rote Kreuze beenden die Verbindung bei Kommata, Satzende oder fehlenden verbindenden Ausdrücken.

2.4.1.2. Behandlung von Auflistungen

Eine Auflistung ist gegeben, wenn mindestens zwei Entitäten der gleichen Suchkategorie (Enzym, Organismus, Ligand, Gewebe und Einheit) durch Junktoren wie „and“ oder „or“ zwischen den letzten Entitäten und mit beliebig vielen Kommata davor verknüpft sind.

Mit der direkten Verbindung können Auflistungen beliebiger Länge erfasst werden. Abbildung 12 zeigt den schematischen Ablauf im Fall solch einer Verbindung. Sowohl ATP als auch GTP werden mit dem kinetischen Ausdrück und der Konzentration verbunden.

Abbildung 12: Schema der inhaltlichen direkten Verbindung anhand eines Beispielsatzes mit einer Auflistung; identifizierte Entitäten sind mit einem grünen Kreis unterlegt; blau unterlegte Entitäten sind Teil der Auflistung; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren verbindende Ausdrücke; rote Kreuze beenden die Verbindung.

Ein Fall mit mehreren Auflistungen in einem Satz ist in Abbildung 13 dargestellt. Hier wird das erste Mitglied der ersten Auflistung (ATP) mit dem ersten Mitglied der zweiten Auflistung (3 M) verbunden. Ebenso wird mit den zweiten Mitgliedern der beiden Auflistungen verfahren (GTP und 2 M).

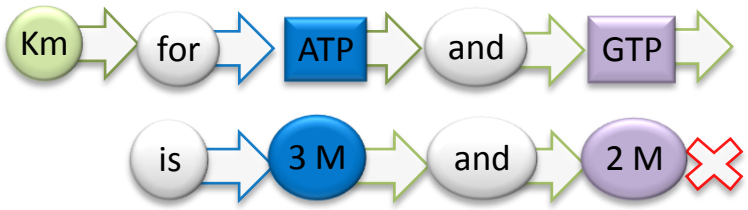

Abbildung 13: Schema der inhaltlichen direkten Verbindung anhand eines Beispielsatzes mit zwei Auflistung; identifizierte Entitäten sind mit einem grünen Kreis unterlegt; blau und violett unterlegte Entitäten sind Teil der Auflistung; die jeweilige Verbindung von den Werten in den Listen untereinander ist durch die gleiche farbliche Unterlegung der Entitäten verdeutlicht; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren zusätzlich verbindende Ausdrücke; rote Kreuze beenden die Verbindung.

Auflistungen von Einheiten wie in Abbildung 14 werden getrennt behandelt. Hier wird von dem letzten Mitglied in der Auflistung die Einheit an die anderen Mitglieder weitergegeben, so dass z.B. „Km“ und „ATP“ nicht mit „5“ sondern mit „5 M“ verbunden werden.

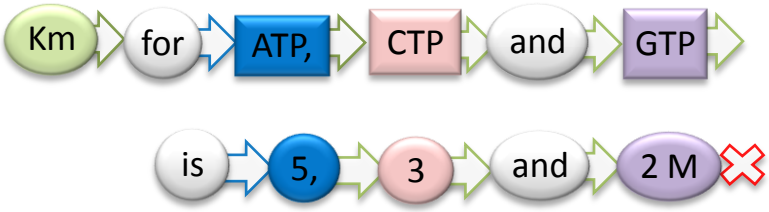

Abbildung 14: Schema der inhaltlichen direkten Verbindung anhand eines Beispielsatzes mit zwei Auflistung für den Sonderfall der Einheiten; identifizierte Entitäten sind mit einem grünen Kreis unterlegt; blau, rosa und violett unterlegte Entitäten sind Teil der Auflistung; die jeweilige Verbindung von den Werten in den Listen untereinander ist durch die gleiche farbliche Unterlegung der Entitäten verdeutlicht; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren zusätzlich verbindende Ausdrücke; rote Kreuze beenden die Verbindung.

2.4.2. Indirekte inhaltliche Verbindung

Sollte eine Kategorie durch die direkte Verbindung nicht gefüllt werden, wird versucht eine indirekte Verbindung vorzunehmen.

Bei der indirekten Verbindung wird geprüft, ob in dem zu untersuchenden Satz nur eine Entität der fehlenden Kategorie enthalten ist. In dem in Abbildung 12 dargestellten Beispiel wird das Enzym „ADH“ inhaltlich indirekt mit „Km“ verbunden.

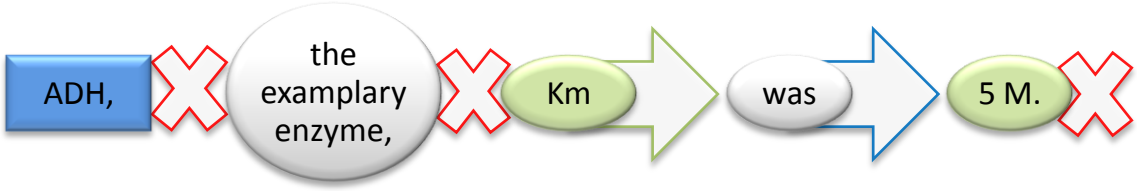

Abbildung 15: Schema der inhaltlichen indirekten Verbindung anhand eines Beispielsatzes; identifizierte Entitäten sind grün oder blau unterlegt; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren verbindende Ausdrücke; rote Kreuze beenden die Verbindung bei Kommata, Satzende oder fehlenden verbindenden Ausdrücken; die blau unterlegte Entität wurde als einzige Entität dieser Kategorie der inhaltlichen Verbindung hinzugefügt ohne direkte Verbindung hinzugefügt.

Konnte die Kategorie nicht auf der Satzebene gefüllt werden, wird geprüft, ob in dem gesamten Abstract nur eine Entität der gesuchten Kategorie enthalten ist. Abschließend wird der Titel auf ein Mitglied diese Kategorie hin untersucht.

2.5. Anpassung und Ergänzung der Ergebnisse

Um die extrahierten Werte miteinander vergleichen zu können, werden sie auf eine einheitliche Größe normiert, z.B. mM bei Konzentrationsangaben.

In Fällen, in denen einem kinetischen Wert zwar ein Enzymname zugeordnet werden konnte, aber die EC-Nummer fehlt, wird diese automatisch ergänzt. Entsprechend wird verfahren, wenn eine EC-Nummer vorhanden ist, aber der Name des Enzyms fehlt.

2.6. Ergebnisdarstellung

Die erhaltenen Ergebnisse werden in einer Datenbank gespeichert und sind dem Nutzer über ein Onlineinterface unter http://134.169.106.54/bth/kid/index.php zugänglich. Zur Erstellung dieses Interfaces wurde auf das Content Management System Joomla [25] zurückgegriffen.

3. Ergebnisse

3.1. Rechenzeit

Das Herunterladen der PubMed-Onlinebibliothek benötigt ca. eine Woche.

Das Extrahieren der kinetischen Informationen aus den Abstracts der PubMed mit 16,2 GB dauert ca. 18 Stunden auf einem Rechner mit einem Turion X2 Prozessor mit 1,6 GHz und beanspruchte 0,3 GB Arbeitsspeicher. Dies entspricht einer durchschnittlichen Rechenzeit von 4 Millisekunden pro Abstract.

3.2. Umfang der verwendeten Daten

3.2.1. Umfang der PubMed

Der Datensatz umfasst 16.953.021 Abstracts der PubMed bis Mai 2007.

3.2.2. Umfang der Lexika

Der Umfang der einzelnen Lexika ist nach Kategorien und absteigend nach ihrem Umfang in den nachfolgenden Tabellen (siehe Tabelle 3 bis Tabelle 5) dargestellt (vergleiche Abschnitt 2).

Tabelle 3: Anzahl der Einträge in den verwendeten Lexika

Suchkategorie	Anzahl der Einträge
Liganden	666.563
Organismen	506.698
Enzymnamen	52.179
EC-Nummern	9.035

Suchkategorie	Anzahl der Einträge
Gewebe	8.512
Ausdrücke für K_M	321
Ausdrücke für K_i	77
Ausdrücke für k_{cat}	76
Ausdrücke für K_d	76
Ausdrücke für IC_{50}	75
Ausdrücke für V_{max}	74
Ausdrücke für n_H	69
Ausdrücke für $t_{1/2}$	64
Ausdrücke für k_{cat}/K_M	42
Ausdrücke für pI	31
Ausdrücke für K_a	28
Ausdrücke für $S_{0.5}$	22
Ausdrücke für spezifische Aktivität	14
Ausdrücke für pH	12
Ausdrücke für Temperatur	6
Einheiten für V_{max}	325
Einheiten für spezifische Aktivität	176
Einheiten für Konzentrationen	116
Einheiten für k_{cat}/K_M	71
Einheiten für $t_{1/2}$	31
Einheiten für k_{cat}	24
Einheiten für K_a	18

Tabelle 4: Anzahl der Einträge in den Ausschlusslexika

Suchkategorie	Anzahl der Einträge
Liganden	472
Enzymnamen	186
Organismen	85
Gewebe	40

Tabelle 5: Hilfslexika für die inhaltliche Verbindung von Entitäten

Verbindende Ausdrücke	2.068	
Markierende Ausdrücke zur Entfernung von Liganden	14	
Markierende Ausdrücke zur Entfernung von Zahlen	Vor der Zahl: 71	Hinter der Zahl: 72
Markierende Ausdrücke zum Ignorieren der Kommata	4	

3.3. Verschiedene Schreibweisen von Zahlen

Die verschiedenen Schreibweisen von Zahlen, die von dem Programm erkannt werden, sind in der nachfolgenden Tabelle exemplarisch zusammengefasst.

Tabelle 6: Beispiele für identifizierbare Formulierungen von Zahlen*

2	2 +/- 3 x 10 (4)	2 up to 3 X 10 (-4)
-2	2 plus/minus 3 x 10 (-4)	2 up to 3 X 10 (4)
2 – 3	2 plus/minus 3 x 10 (4)	2 upto 3 X 10 (-4)
2 to 3	2 plus or minus 3 x 10(-4)	2 upto 3 X 10 (4)
2 upto 3	2 plus or minus 3 x 10(4)	2 - 3 times 10 (-4)
2 up to 3	2 +/- 3 X 10 (-4)	2 - 3 times 10 (4)
2 +/- 3	2 +/- 3 X 10 (4)	2 to 3 times 10 (-4)
2 plus/minus 3	2 plus/minus 3 X 10 (-4)	2 to 3 times 10 (4)
2 plus or minus 3	2 plus/minus 3 X 10 (4)	2 up to 3 times 10 (-4)
2 000	2 plus or minus 3 X 10 (-4)	2 up to 3 times 10 (4)
2,000	2 up to 3 x 10 (-4)	2 upto 3 times 10 (-4)
3 x 10 (-4)	2 up to 3 x 10 (4)	2 upto 3 times 10 (4)
3 x 10 (4)	2 upto 3 x 10 (-4)	2 plus or minus 3 . 10 (4)
3 X 10 (-4)	2 upto 3 x 10 (4)	2 plus or minus 3 . 10 (-4)
3 X 10 (4)	2 - 3 X 10 (-4)	2 plus/minus 3 . 10 (4)
3 times 10 (-4)	2 - 3 X 10 (4)	2 plus/minus 3 . 10 (-4)
3 times 10 (4)	2 to 3 X 10 (-4)	2 +/- 3 . 10 (4)
3 . 10 (-4)	2 to 3 X 10 (4)	2 +/- 3 . 10 (-4)
3 . 10 (4)	2 - 3 x 10 (-4)	2 plus or minus 3 times 10 (4)
3 x (10-4)	2 - 3 x 10 (4)	2 plus or minus 3 times 10 (-4)
3 X (10-4)	2 - 3 x 10 (4)	2 plus/minus 3 times 10 (4)
3 times (10-4)	2 to 3 x 10 (-4)	2 plus/minus 3 times 10 (-4)
3 x 10-4	2 to 3 x 10 (4)	2 +/- 3 times 10 (4)

3 X 10-4	2 up to 3 x 10 (-4)	2 +/- 3 times 10 (-4)
3 times 10-4	2 up to 3 x 10 (4)	2 plus or minus 3 X 10 (4)
10(-4)	10-4	2 +/- 3 x 10 (-4)
10(4)		

*Anmerkung zu Tabelle 6: jede dargestellte Zahl (2, 3, 4) kann auch eine Dezimalzahl mit beliebig vielen Nachkommastellen (z.B. 2.5, 2.56, …) sein. Leerzeichen sind optional.

3.4. Darstellung der Ergebnisse

Über eine Suchmaske im Interface ist der Nutzer in der Lage den Inhalt der Datenbank nach seinen Wünschen zu filtern, wobei die Ergebnisse anschließend in einer tabellarischen Form präsentiert werden (siehe Abbildung 16).

PubMedID	Expression	Value	Enzyme	Ligand
360736 2	Km	2.5 x 10(-2) M,	alpha-glucosidase	maltose
365220 0	Km	1.66 x 10(-2) M,	alpha-glucosidase	maltose
2689637 1	Km	2.5 mM,	acid alpha-glucosidase	maltose,
2689637 2	Km	28.5 mM	acid alpha-glucosidase	isomaltose
3891151 0	(Km,	0.8 mmol/l)	alpha-glucosidase	maltose
6362728 0	Km	0.7 mM)	alpha-glucosidase	maltose
6394601 0	Km	6.3 mM	acid alpha-glucosidase	maltose
7052147 0	Km	5 mM,	alpha-glucosidase	maltose
7052147 1	Km	6 mM,	alpha-glucosidase	maltose
16349368 1	K(m)	3.0, mM,	alpha-Glucosidase	maltose,
16667773 1	apparent K(m)	33 millimolar	alpha-Glucosidase	maltose

Abbildung 16: Tabelle der Suchergebnisse mit Verknüpfungen (unterstrichen und rot unterlegt) zu einer ausführlicheren Präsentation der Ergebnisse.

Über die PubMed ID gelangt der Nutzer zu einer ausführlicheren Darstellung der Ergebnisse (siehe Abbildung 17). Diese umfasst sowohl eine Abbildung des Abstract-Textes, in dem die Ergebnisse eingefärbt werden, als auch eine tabellarische Darstellung der Ergebnisse.


PubMed ID: 365220 Result ID: 0

Purification and characterization of an **alpha-glucosidase** from Saccharomyces carlsbergensis.

RB Needleman; HJ Federoff; TR Eccleshall; B Buchferer; J Marmur;

Biochemistry; Vol. 17; Oct. 1978; 4657-61;

alpha-Glucosidase (EC **3.2.1.20)** was purified to homogeneity from logarithmically growing **cells** of **Saccharomyces carlsbergensis.** The purification involved the following steps: (a) ammonium sulfate fractionation; (b) Sephadex G-100 chromatography; (c) DEAE-cellulose chromatography; and (d) hydroxylapatite chromatography. This procedure gave a preparation judged to be greater than 98% pure by Na-DodSO4-polyacrylamide gel electrophoresis. The enzyme was shown to be a monomer of 63 000 daltons by gel filtration on Sephacryl S-200 under native conditions and by polyacrylamide gel electrophoresis under denaturing conditions. The ***Km*** values of the enzyme for the substrates **maltose** and p-nitrophenyl alpha-D-glucoside were found to be ***1.66 X 10(-2)*** and 3.1 X 10(-4) M, respectively. The corresponding Vmax value for maltose was 44.8 X 10(-6) mol min(-1) mg(-1) and that for p-nitrophenyl alpha-D-glucoside was 134 X 10(-6) mol min-1 mg-1. The pH optimum for the purified enzyme was found to be between pH 6.7 and 6.8. The enzyme has an absolute anomeric specificity for alpha-glycosidic linkages and appears to recognize a glucosyl residue in alpha linkage on the nonreducing end of its substrate. For the strain used in this study, which carries the MAL 6 locus, only a single form of the enzyme was detected.

Expression for Km:	*Km*
EC number:	3.2.1.20)
Location:	cells
Organism:	Saccharomyces carlsbergensis
Conzentration:	*1.66 x 10(-2) M,*
Enzyme or receptor:	alpha-glucosidase
Ligand:	maltose

Validate or edit

Abbildung 17: Ausführliche Darstellung der Ergebnisse.

Durch das Einfärben der extrahierten Werte im Text soll dem Nutzer das Erfassen und Überprüfen der inhaltlichen Verbindungen der Ergebnisse erleichtert werden.

Zusätzlich hat der Nutzer die Möglichkeit die Ergebnisse zu überprüfen und Verbesserungen vorzuschlagen.

3.5. Quantitative Analyse der Ergebnisse

3.5.1. Gesamtanzahl extrahierter Werte

Die nachfolgende Tabelle gibt einen Überblick über die Gesamtanzahl der extrahierten Entitäten in der jeweiligen Kategorie.

Tabelle 7: Gesamtanzahl der extrahierten Werte unterteilt nach ihren Kategorien

Kategorie	Gesamtzahl der extrahierten Werte
kinetischen Ausdrücke	509.153
Liganden	338.706
Organismen	332.579
Lokalisation	285.453
Enzyme	267.356
Zahlenwerte	244.124
EC Nummer	77.001
Temperatur	10.945
pH	10.112

In Abbildung 18 und Abbildung 19 werden die Mengen der extrahieren Entitäten in den verschieden Kategorien bezogen auf den verbunden kinetischen Ausdruck dargestellt. Die Gesamtanzahl gekennzeichnet die Anzahl der extrahierten kinetischen Ausdrücke.

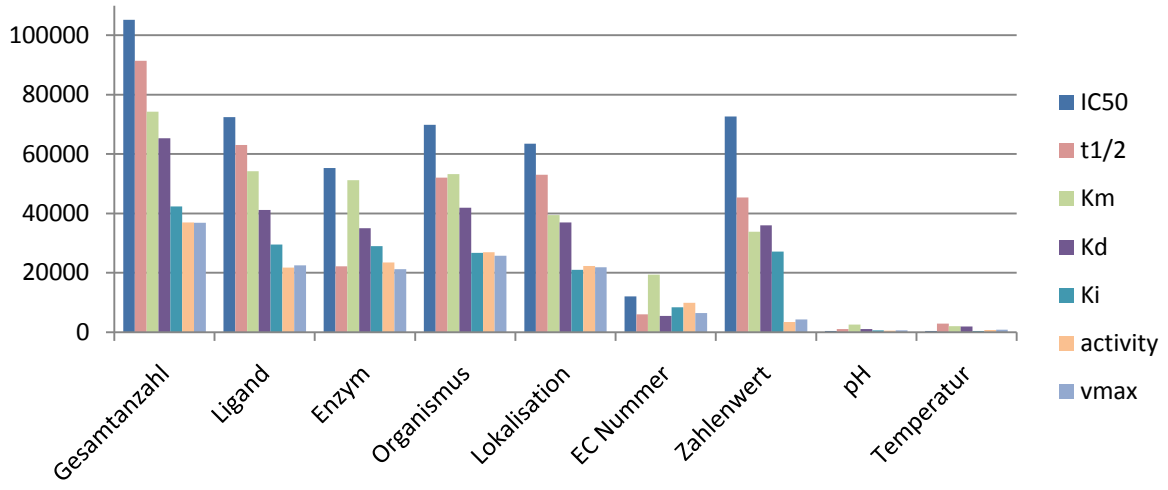

Abbildung 18: Abbildung 18Gesamtanzahl der extrahierten Werte unterschieden nach verbundenem kinetischem Wert und Kategorie Teil 1 (siehe Anhang Punkt 7.1.1).

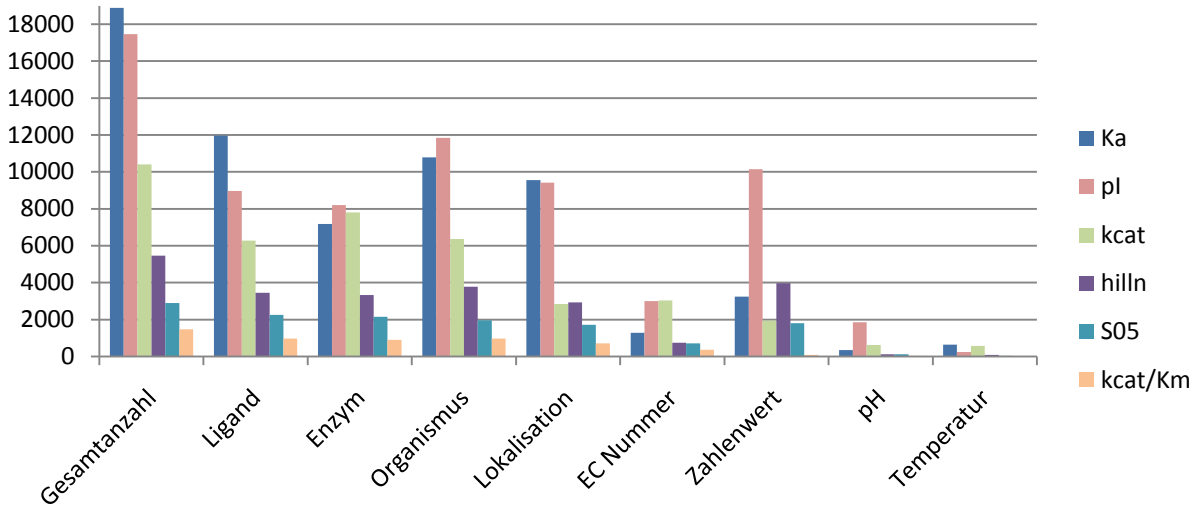

Abbildung 19: Gesamtanzahl der extrahierten Werte unterschieden nach verbundenem kinetischem Wert und Kategorie Teil 2 (siehe Anhang Punkt 7.1.1).

Aus Abbildung 18 ist ersichtlich, dass die meisten Informationen zu den kinetischen Kategorien IC_{50} (105.240 Einträge), $t_{1/2}$ (91.430 Einträge), K_M (74.253 Einträge) und K_d (65.401 Einträge) gefunden wurden. Für k_{cat}/K_M wurden mit 1.469 Einträgen die wenigsten Informationen extrahiert (siehe Abbildung 19).

Bei der Betrachtung der Gesamtanzahlen in den einzelnen Kategorien fällt auf, dass für alle kinetischen Ausdrücke der pH-Wert und die Temperatur am wenigsten extrahierte Entitäten aufweisen (vergleiche Abbildung 18 und Abbildung 19). Auch EC Nummern wurden in vergleichsweise kleinen Mengen zugewiesen.

In Tabelle 8 sind die Mengen der Abstracts bezogen auf den identifizierten kinetischen Ausdruck aufgeführt.

Tabelle 8: Anzahl der Abstracts bezogen auf den identifizierten kinetischen Ausdruck

Kinetischer Ausdruck	Anzahl der Abstracts
$t_{1/2}$	57.658
IC_{50}	54.938
K_M	45.896
K_d	40.244
spezifische Aktivität	27.013
V_{max}	23.985
K_i	22.805

Kinetischer Ausdruck	Anzahl der Abstracts
pI	13.313
K_a	10.502
k_{cat}	7.325
n_H	4.046
k_{cat}/K_M	3.587
$S_{0.5}$	1.571

Die meisten Abstracts enthalten Informationen zu $t_{1/2}$, IC_{50}, K_M und der spezifischen Aktivität. Die geringste Anzahl an Abstracts wurde zu K_{cat}/K_M und $S_{0.5}$ gefunden.

3.5.2. Kombinationen extrahierter Werte

In Abbildung 20 bis Abbildung 23 sind die prozentualen Verteilungen von Kombinationen verschiedener Kategorien dargestellt. Für Abbildung 20 und Abbildung 21 sind die Prozentzahlen auf die Anzahl der verschiedenen extrahierten kinetischen Ausdrücke bezogen.

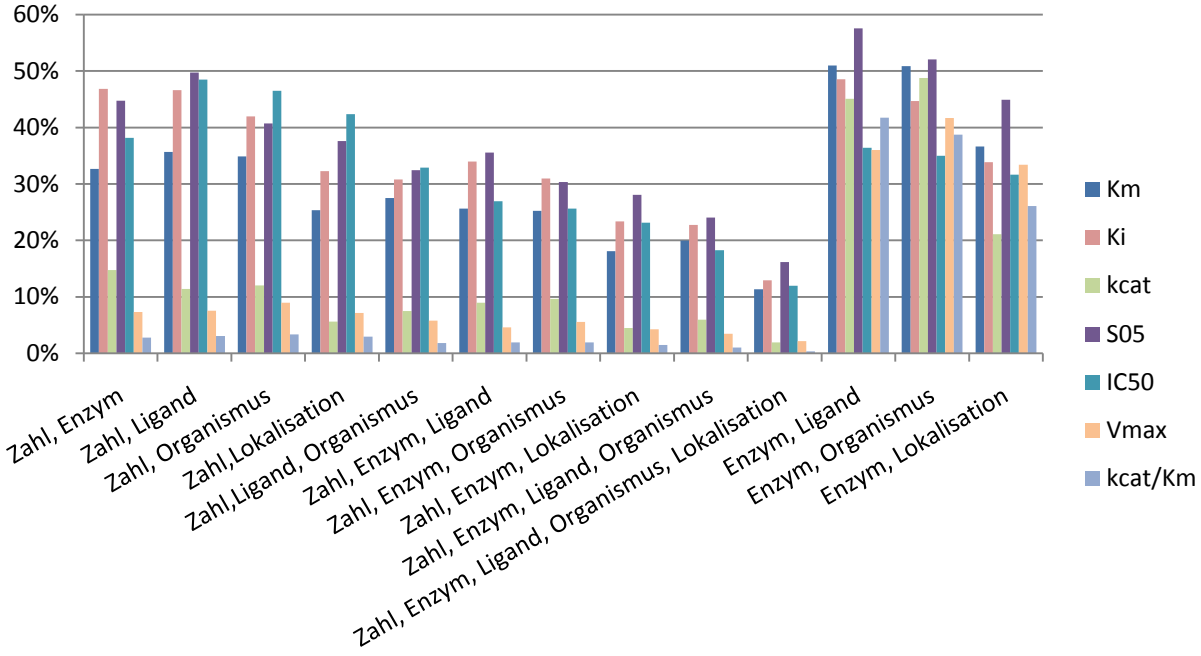

Abbildung 20: Prozentuale Verteilung der Kombinationen verschiedener Kategorien bezogen auf die Anzahl der jeweiligen kinetischen Kategorie Teil 1. Zahl stellt eine Abkürzung für einen Zahlenwert dar; Enzym ist eine Abkürzung für Enzyminformation und umfasst sowohl Enzymnamen als auch EC Nummern (siehe Anhang Punkt 7.1.2).

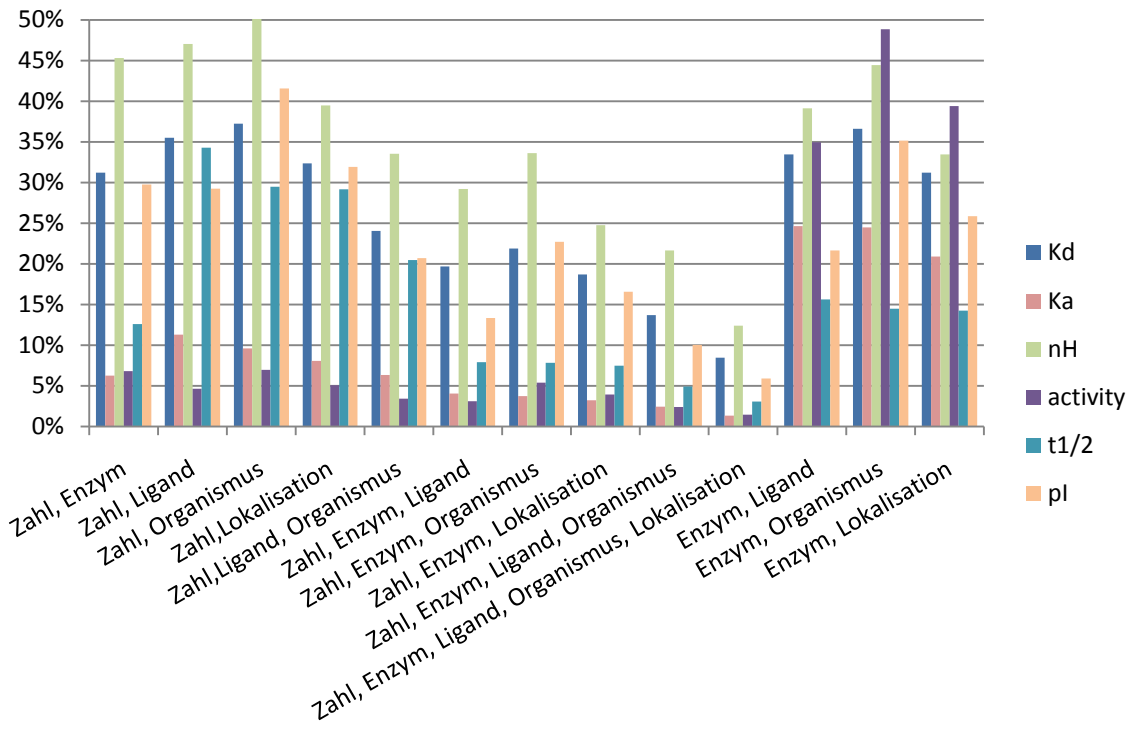

Abbildung 21: Prozentuale Verteilung der Kombinationen verschiedener Kategorien bezogen auf die Anzahl der jeweiligen kinetischen Kategorie Teil 2. Zahl stellt eine Abkürzung für einen Zahlenwert dar; Enzym ist eine Abkürzung für Enzyminformation und umfasst sowohl Enzymnamen als auch EC Nummern (siehe Anhang Punkt 7.1.2).

Anhand der obigen Graphiken ist ersichtlich, dass die Kombinationen, die einen Zahlenwert enthalten in kleineren prozentualen Mengen vorliegen, als Kombinationen ohne Zahlenwert. Zusätzlich ist zu erkennen, dass die Menge der Kombination mit zunehmender Komplexität abnimmt.

In den folgenden Abbildungen (Abbildung 22 und Abbildung 23) sind die prozentualen Verteilungen der Kombinationen bezogen auf die Anzahl des jeweiligen kinetischen Wertes mit Zahlenwert dargestellt.

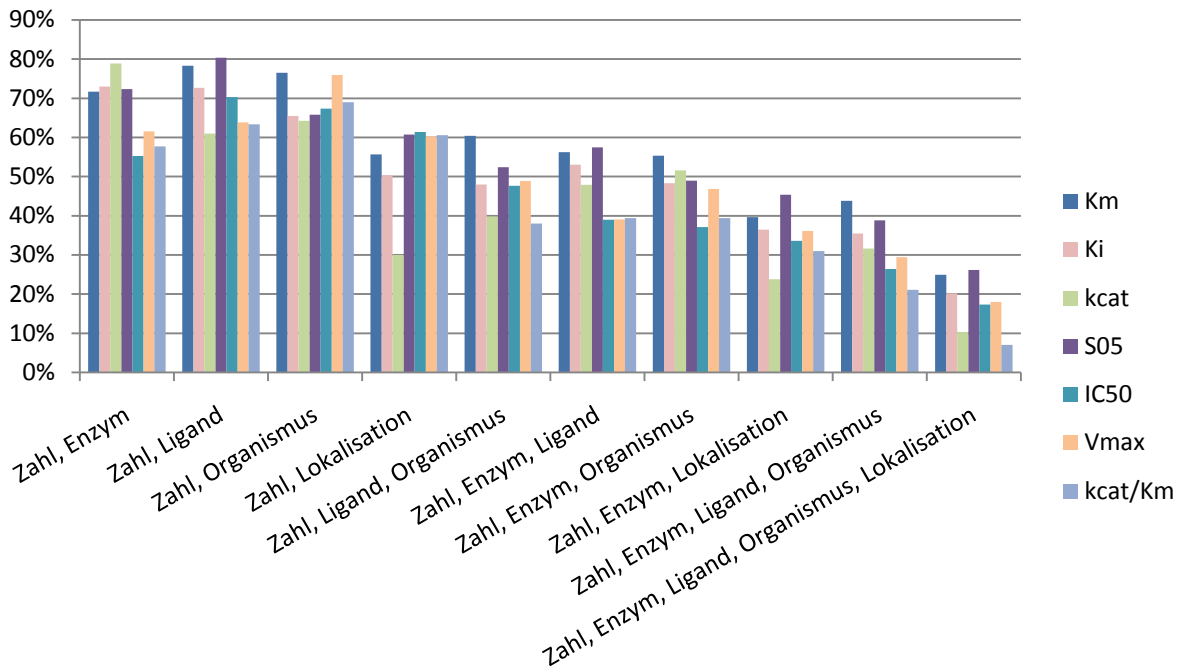

Abbildung 22: Prozentuale Verteilung der Kombinationen verschiedener Kategorien bezogen auf die Anzahl der jeweiligen kinetischen Kategorie mit einem zugewiesen Zahlenwert Teil 1. Zahl stellt eine Abkürzung für einen Zahlenwert dar; Enzym ist eine Abkürzung für Enzyminformation und umfasst sowohl Enzymnamen als auch EC Nummern (siehe Anhang Punkt 7.1.2).

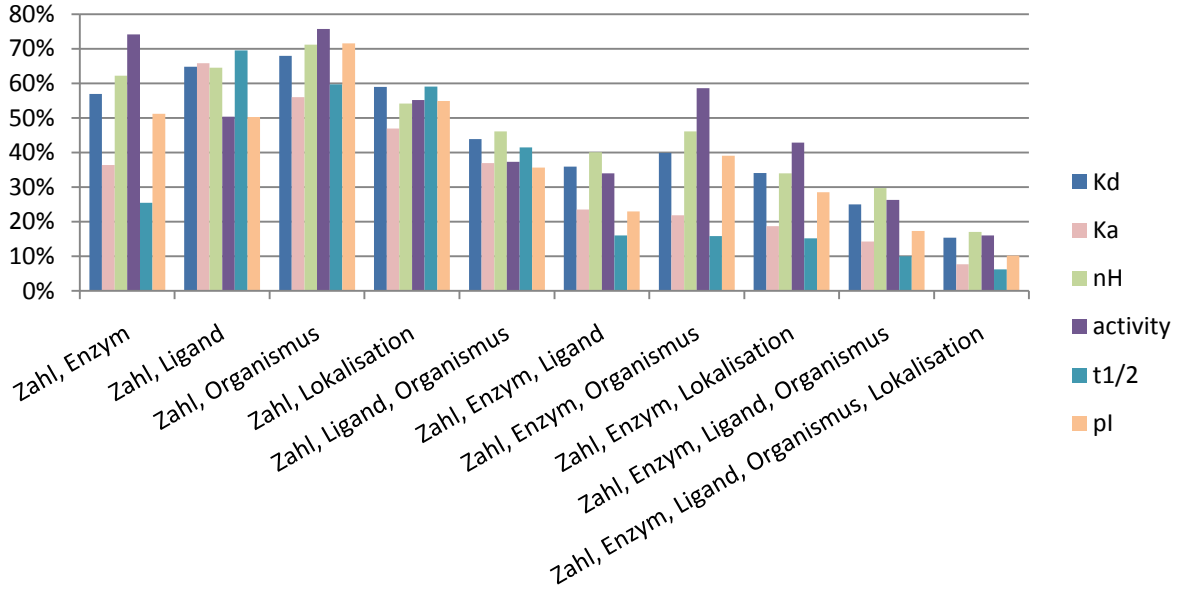

Abbildung 23: Prozentuale Verteilung der Kombinationen verschiedener Kategorien bezogen auf die Anzahl der jeweiligen kinetischen Kategorie mit einem zugewiesen Zahlenwert Teil 2. Zahl stellt eine Abkürzung für einen Zahlenwert dar; Enzym ist eine Abkürzung für Enzyminformation und umfasst sowohl Enzymnamen als auch EC Nummern (siehe Anhang Punkt 7.1.2).

Auffällig ist, dass für K_a und $t_{1/2}$ vergleichsweise wenige Enzyminformationen verknüpft werden konnten (siehe Abbildung 23).

3.5.3. Art der Verbindung

In den folgenden Abbildungen sind die Häufigkeiten der verschiedenen Arten der Werteverbindung beispielhaft für IC_{50} und K_M dargestellt (siehe Abbildung 24 und Abbildung 25).

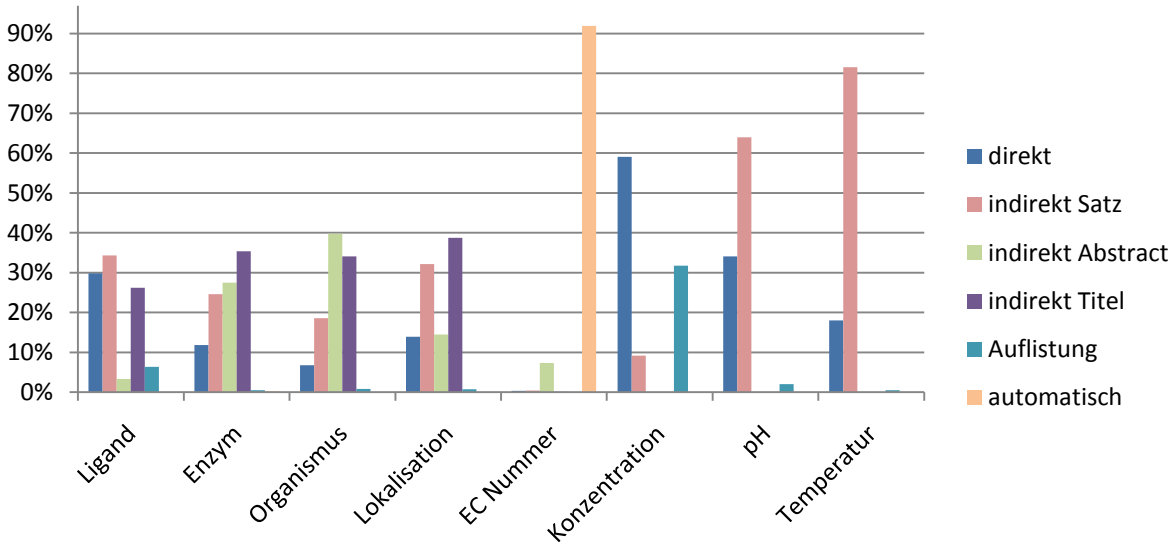

Abbildung 24: Darstellung der Häufigkeiten verschiedener Arten der Werteverbindung für IC_{50} (siehe Anhang Punkt 7.1.3).

In der graphischen Auftragung für IC_{50} lässt sich erkennen, dass die meisten Liganden direkt (30%) oder indirekt im Satz (34%) verbunden werden. 26% der Liganden wurden im Titel gefunden. Die Suche über den Abstract brachte vergleichsweise wenige Liganden (3%). 6% der Liganden konnten aus Auflistungen extrahiert werden.

Bei den Enzymen wurde der Großteil über die Suche im Titel (35%) oder Abstract (27%) gefunden. 25% der Enzyme konnten indirekt aus dem Satz extrahiert werden. Eine direkte Verbindung (12%) oder die Extraktion aus Auflistungen (0,5%) fand vergleichsweise selten statt. Der Anteil der automatisch zugewiesenen Enzymnamen liegt in diesem Fall unter 0,3%.

Für Organismen brachte die Suche über den Abstract (40%) und über den Titel (34%) die meisten Ergebnisse. Nur in wenigen Fällen fand eine direkte Verbindung oder eine Extraktion aus einer Auflistung statt.

Der Großteil der Lokalisationen wurde mittels der Suche über den Titel (39%) und indirekt über den Satz (32%) verbunden. Eine direkte Verbindung oder die Suche über den Abstract brachten weniger Ergebnisse.

Der Großteil der EC Nummern wurde mit 91% automatisch annotiert. Alle anderen Arten der Verbindung waren vergleichsweise selten.

59% der zugewiesenen Zahlenwerte wurden über eine direkte Verbindung oder eine Auflistung (32%) zugewiesen. 9% der Zahlen wurden indirekt über den Satz verbunden. Alle anderen Arten der Verbindungen waren weniger oft vertreten.

Informationen zu pH-Werten und Temperaturen wurden großteilig indirekt im Satz (64% und 81%) oder direkt (34% und 18%) zugewiesen.

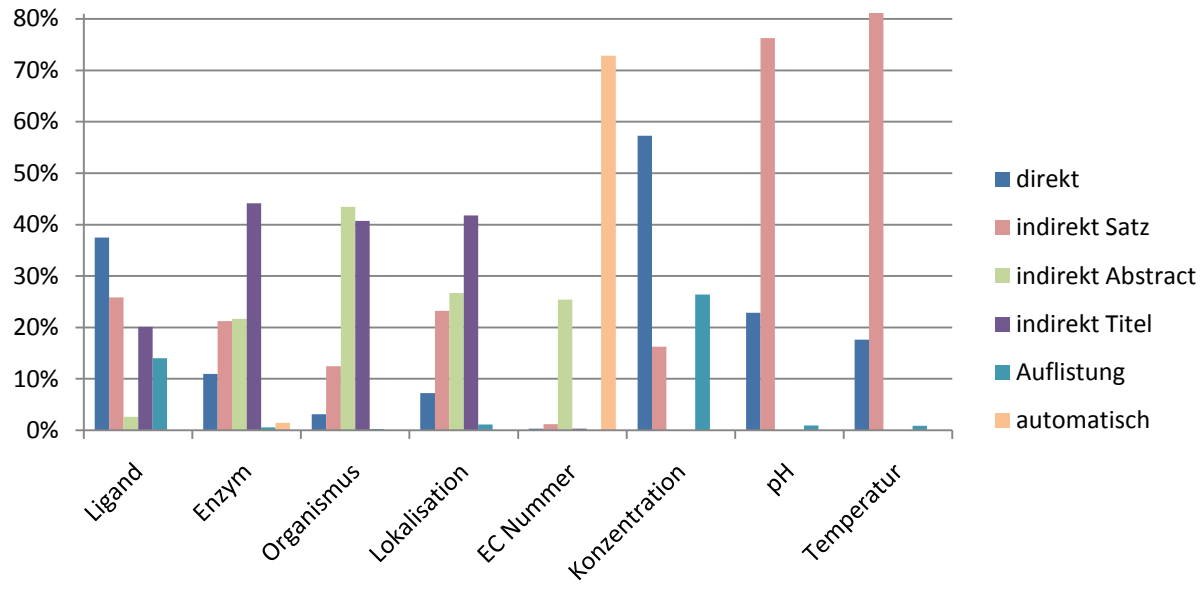

Abbildung 25: Darstellung der Häufigkeiten verschiedener Arten der Werteverbindung für K_M (siehe Anhang Punkt 7.1.3).

Wie in Abbildung 25 ersichtlich wird, wird der Großteil der Liganden (37%) direkt verbunden, wohingegen 26% der Liganden indirekt im Satz zugewiesen wird. Auch die Suche über den Titel (20%) und die Extraktion aus Auflistungen (14%) wurden verwendet.

Im Fall der Enzyme werden die meisten Ergebnisse mittels Suche über den Titel (44%) erhalten. Die Suche über den Satz (21%) und das Abstract (22%) führt ebenfalls zu Ergebnissen. Nur ein geringer Teil (11%) konnte direkt verbunden werden.

Bei den Organismen wurde die Suche über das Abstract (43%) und über den Titel (41%) am häufigsten angewendet.

Für Lokalisationen wurde überwiegend die Suche über den Titel (42%) und über den Abstract (27%) verwendet. 23% konnten indirekt über den Satz zugewiesen werden.

Für EC Nummern konnten nur vergleichsweise wenige Verbindungen durchgeführt werden, wobei die Suche über den Abstract die meisten Ergebnisse bringt. Der Großteil der Entitäten wird automatisch (72%) annotiert.

Die direkte Verbindung ist bei der Zuweisung eines Zahlenwertes mit 57% am erfolgreichsten. 26% der Werte konnten aus Auflistungen extrahiert werden, wobei in 16% der Fälle die Verbindung indirekt im Satz statt findet.

Für pH-Werte und Temperaturen konnten 76% und 81% indirekt im Satz zugewiesen werden.

3.5.4. Verteilung von extrahierten kinetischen Zahlenwerten

Die Präferenzbereiche der verschiedenen extrahierten Werte sind in den folgenden Abbildungen dargestellt. Auf die Ergebnisse für k_{cat}/K_M wird aufgrund der geringen Menge nicht näher eingegangen.

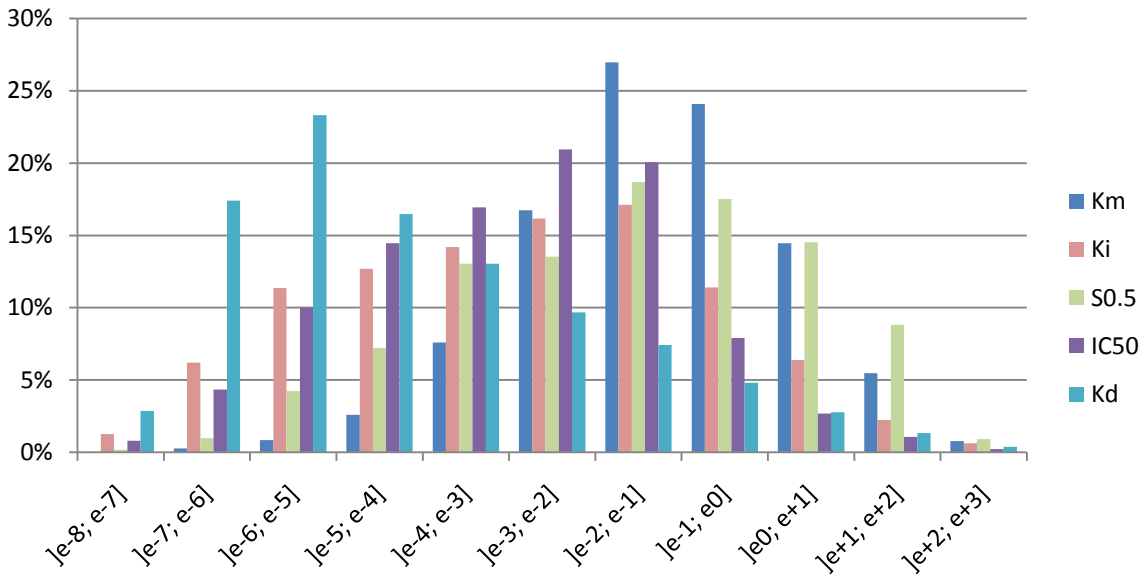

Abbildung 26: Prozentuale Verteilung von K_d und verschiedenen kinetischen Werten [mM] (siehe Anhang Punkt 7.1.4).

In Abbildung 26 zeigen die Werte für K_d und verschiedene kinetische Kategorien eine nahezu glockenartige Verteilung. Werte für K_i, K_M und $S_{0.5}$ zeigen ein Maximum im Bereich von 10^{-2} bis 10^{-1} mM. Die maximale Anzahl der extrahierten IC_{50}-Werte liegt bei 10^{-3} bis 10^{-2} mM und liegt somit in der Nähe der zuvor erwähnten kinetischen Werte. Im Gegensatz dazu liegt das Maximum für K_d bei 10^{-6} bis 10^{-5} mM.

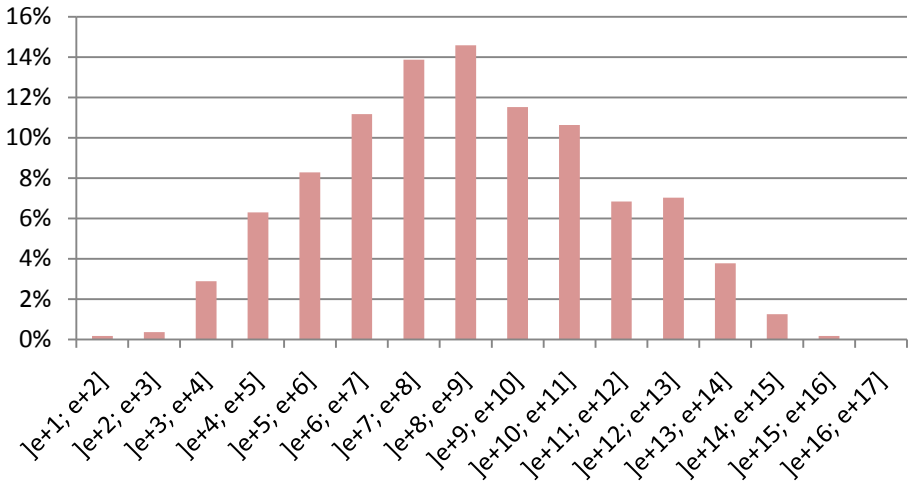

Abbildung 27: Prozentuale Verteilung von K_a-Werten [mM^{-1}] (siehe Anhang Punkt 7.1.4).

Auch in Abbildung 27 zeigt sich für die Verteilung von K_a-Werten ein Maximum im Bereich von 10^{+8} bis 10^{+9} mM^{-1}.

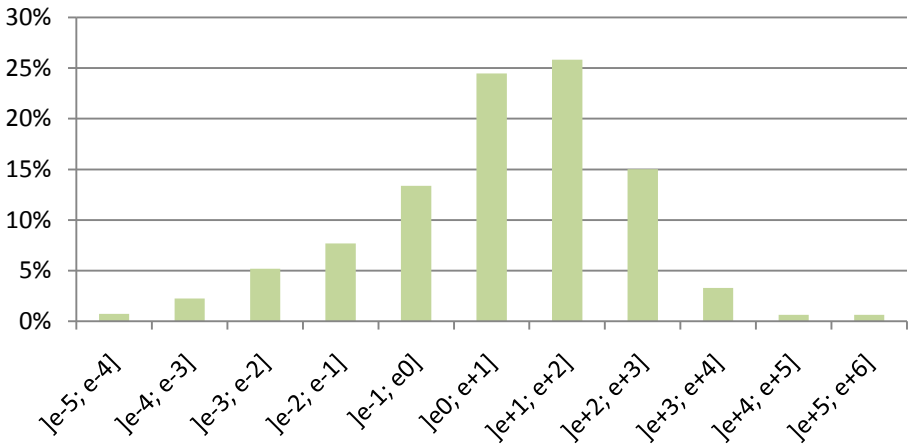

Abbildung 28: Prozentuale Verteilung von k_{cat}-Werten [s^{-1}] (siehe Anhang Punkt 7.1.4).

Für k_{cat}-Werte lässt sich ein Maximum im Bereich von 10^0 bis 10^2 s^{-1} feststellen (siehe Abbildung 28).

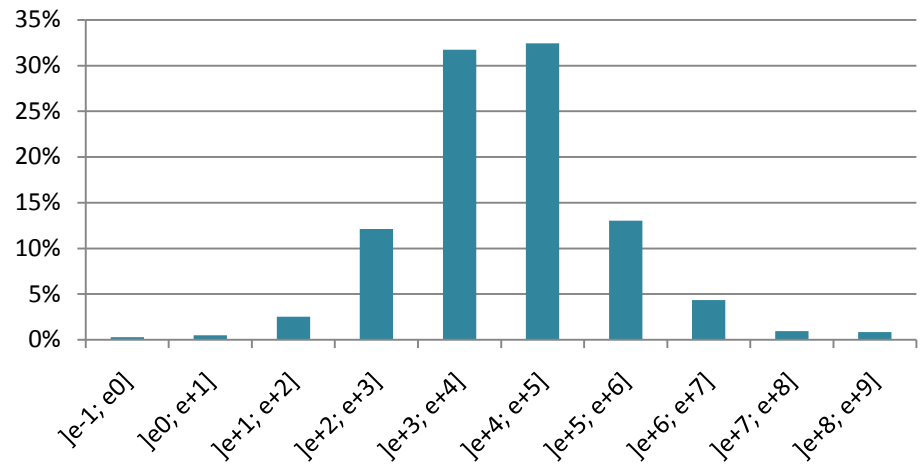

Abbildung 29: Prozentuale Verteilung von $t_{1/2}$-Werten [s] (siehe Anhang Punkt 7.1.4).

Das Maximum der Verteilung für $t_{1/2}$-Werte lässt sich bei 10^3 bis 10^5 s ermitteln (siehe Abbildung 29).

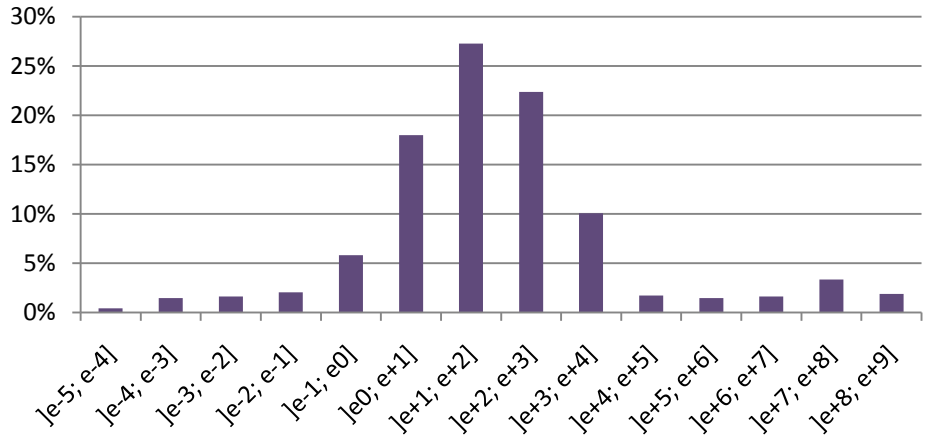

Abbildung 30: Prozentuale Verteilung von Werten für die spezifische Aktivität [U/mg] (siehe Anhang Punkt 7.1.4).

Aus Abbildung 30 wird ersichtlich, dass die meisten Werte für die spezifische Aktivität sich in dem Bereich von 10^1 bis 10^2 Unit/mg befinden.

3.5.5. Verteilung der EC-Klassen

In der nachfolgenden Tabelle 9 sind die Häufigkeiten der EC Nummern in den Ergebnissen und den Lexika dargestellt.

Tabelle 9: Häufigkeit der EC Nummern in den Ergebnissen und Lexika

	Gesamtzahl	Anzahl verschiedener EC Nummern einer Klasse in den Ergebnissen	Anzahl unterschiedlicher EC Nummern einer Klasse in den Lexika
EC Klasse 1	16.576	441	1.192
EC Klasse 2	18.663	470	1.189
EC Klasse 3	33.830	540	1.416
EC Klasse 4	6.317	166	413
EC Klasse 5	955	59	165
EC Klasse 6	2.119	79	143

3.5.6. Häufigkeit und Verteilung extrahierter Organismen

Abbildung 31 zeigt die Verteilung der Organismen in den Ergebnissen bezogen auf verschiedenen Organismengruppen, wobei zu erkennen ist, dass überwiegend Informationen zu Tieren bzw. Säugetieren extrahiert werden.

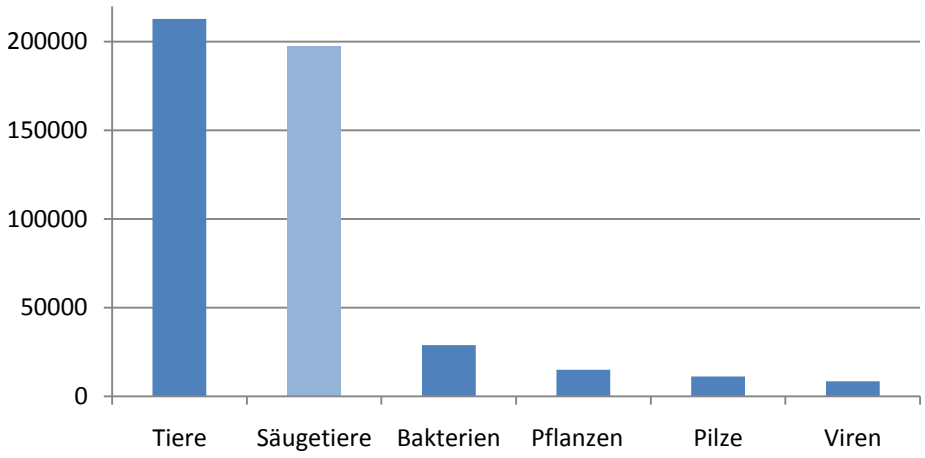

Abbildung 31: Häufigkeiten verschiedener Organismengruppen [blaue Balken] in den Ergebnissen, Säugetiere [hell blauer Balken] sind eine Untergruppe der Tiere(siehe Anhang Punkt 7.1.5).

In Tabelle 10 sind die häufigsten Organismen mit ihrer Anzahl aufgeführt.

Tabelle 10: Anzahlen der häufigsten Organismen

Organismus	Anzahl
Mensch	86.338
Ratte	56.456
Maus	15.368
Kaninchen	1.5368
Rind	11.379
Schwein	9.326
Escherichia coli	9.205
Hefe	4.641

3.5.7. Verteilung der Publikationsdaten

In den folgenden Abbildung 32 und Abbildung 33 ist die Anzahl der Abstracts mit kinetischem Inhalt gegen das Jahr der Veröffentlichung aufgetragen.

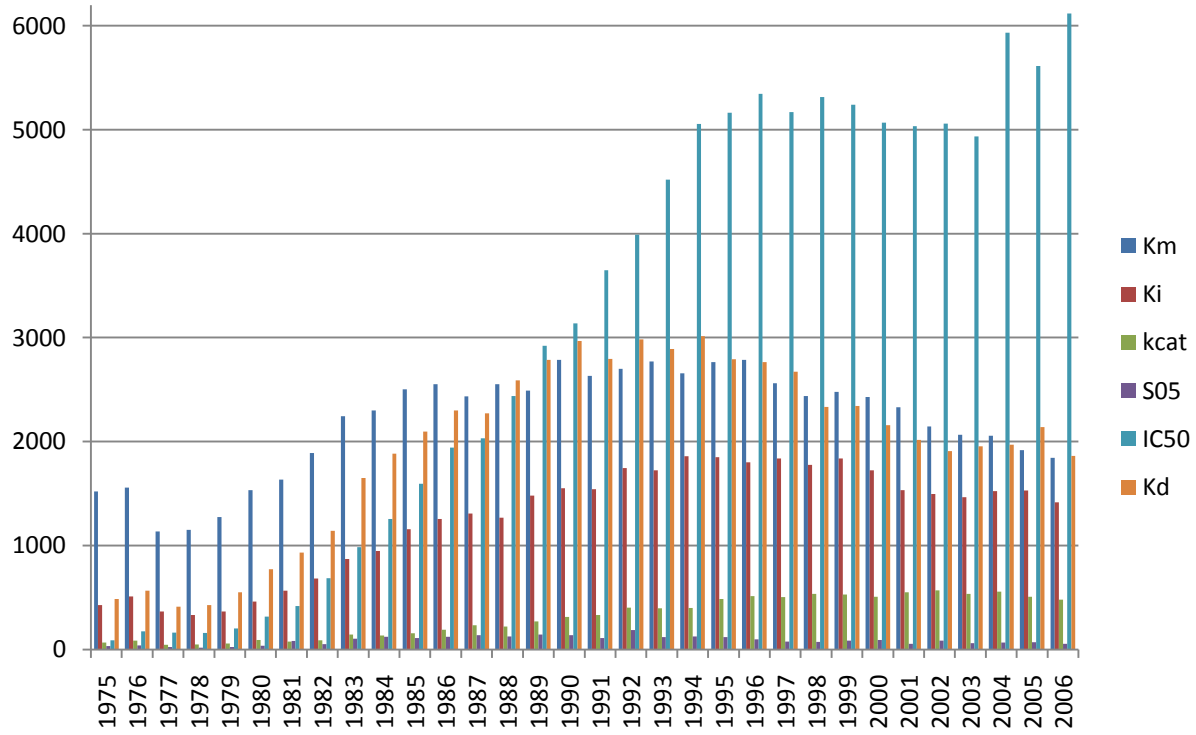

Abbildung 32: Menge an Abstracts bezogen auf den beinhalteten kinetischen Ausdruck und das Jahr der Veröffentlichung Teil 1 (siehe Anhang Punkt 7.1.6).

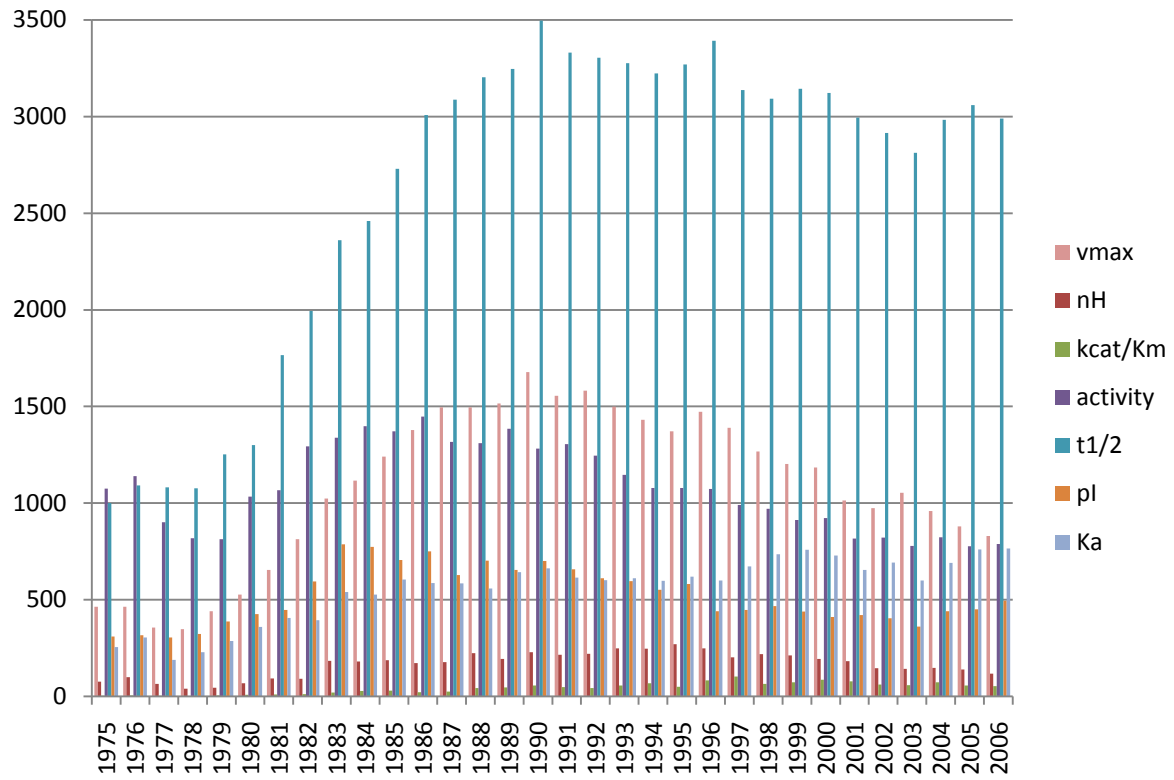

Abbildung 33: Menge an Abstracts bezogen auf den beinhalteten kinetischen Ausdruck und das Jahr der Veröffentlichung Teil 2 (siehe Anhang Punkt 7.1.6).

Aus den obigen Abbildungen lässt sich erkennen, dass die Menge an Abstracts mit kinetischem Inhalt seit den 1980er Jahren zugenommen hat. Auffällig ist, dass die Menge mit Ausdrücken für K_M, K_i, und V_{max} in den 1980ern und 1990ern ihren Höhepunkt erreicht. In den nachfolgenden Jahren ist eine geringfügige Abnahme zu verzeichnen. Die Menge an Veröffentlichungen zur spezifischen Aktivität fällt von 1975 bis 1979 ab. Anschließend steigt sie bis 1984 an und sinkt danach wieder ab. Informationen zu k_{cat} scheinen über den aufgezeigten Zeitraum bis 2005 stetig anzusteigen und fallen dann ab. Für $S_{0.5}$, k_{cat}/K_M und n_H ist keine besondere Hochzeit erkennbar, da die Menge an Informationen zu diesem Wert im Vergleich gering ist. Bei IC_{50} lässt sich ein weitgehend stetiger Anstieg der Informationen in der Literatur bis 2006 erkennen. Die Menge an Informationen zu $t_{1/2}$ steigt ebenfalls seit Anfang der 1980er Jahre an und sinkt von Ende der 1990er Jahre bis Anfang 2000 ab. Die Menge an publizierten Daten zu pI ist in den 1980er und 1990er Jahren leicht erhöht und bleibt danach weitgehend konstant. Im Gegensatz dazu zeigt sich für K_a eine über die Jahre weitgehend gleichbleibende Menge an Publikationen. Für K_d ist ein Anstieg von Anfang 1980er Jahre bis Mitte der 1990er Jahre erkennbar, welcher von einem Abfall gefolgt wird.

In den Abbildung 34 und Abbildung 35 sind die Mengen der Abstracts mit kinetischem Inhalt auf die Gesamtmenge der erschienenen Abstracts bezogen. Bei dieser Auftragung sind für die verschiedenen kinetischen Kategorien ähnliche zeitliche Verteilungen zu erkennen wie in den Abbildung 32 und Abbildung 33, die Maxima erscheinen jedoch ausgeprägter.

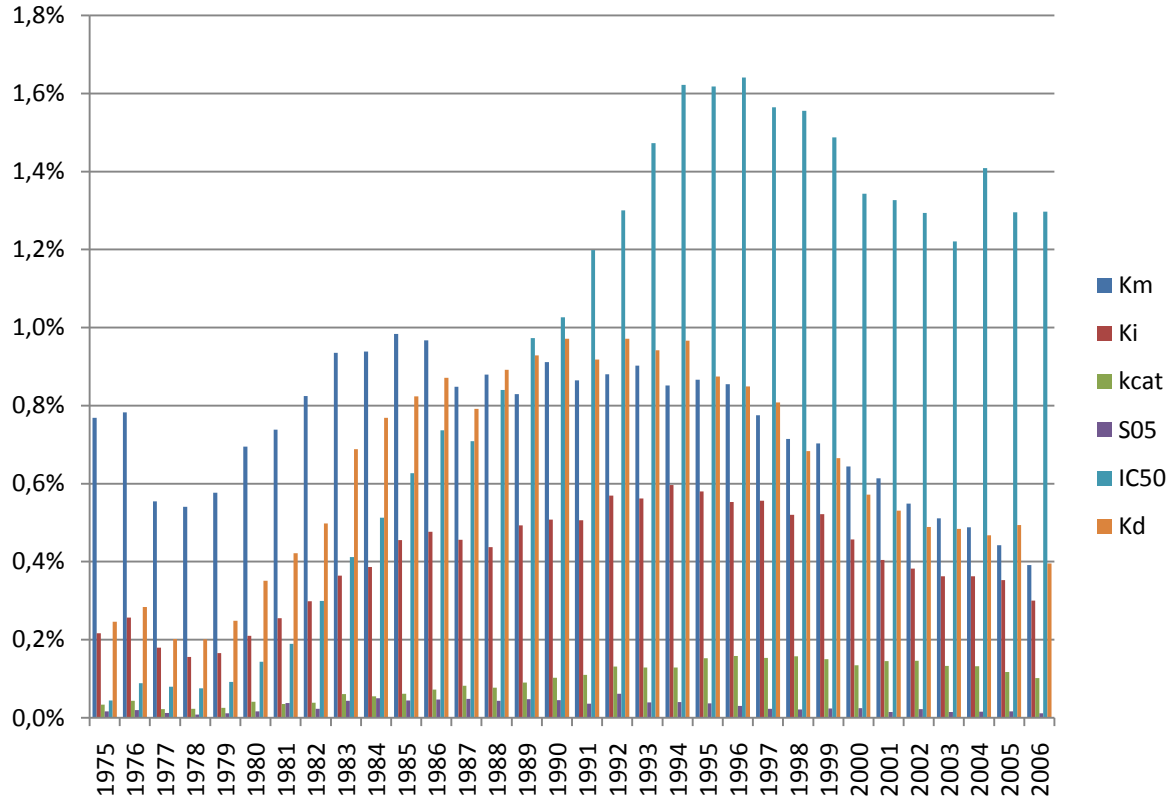

Abbildung 34: Menge an Abstracts bezogen auf den beinhalteten kinetischen Ausdruck und die Menge an veröffentlichten Abstract im Erscheinungsjahr Teil 1 (siehe Anhang Punkt 7.1.6).

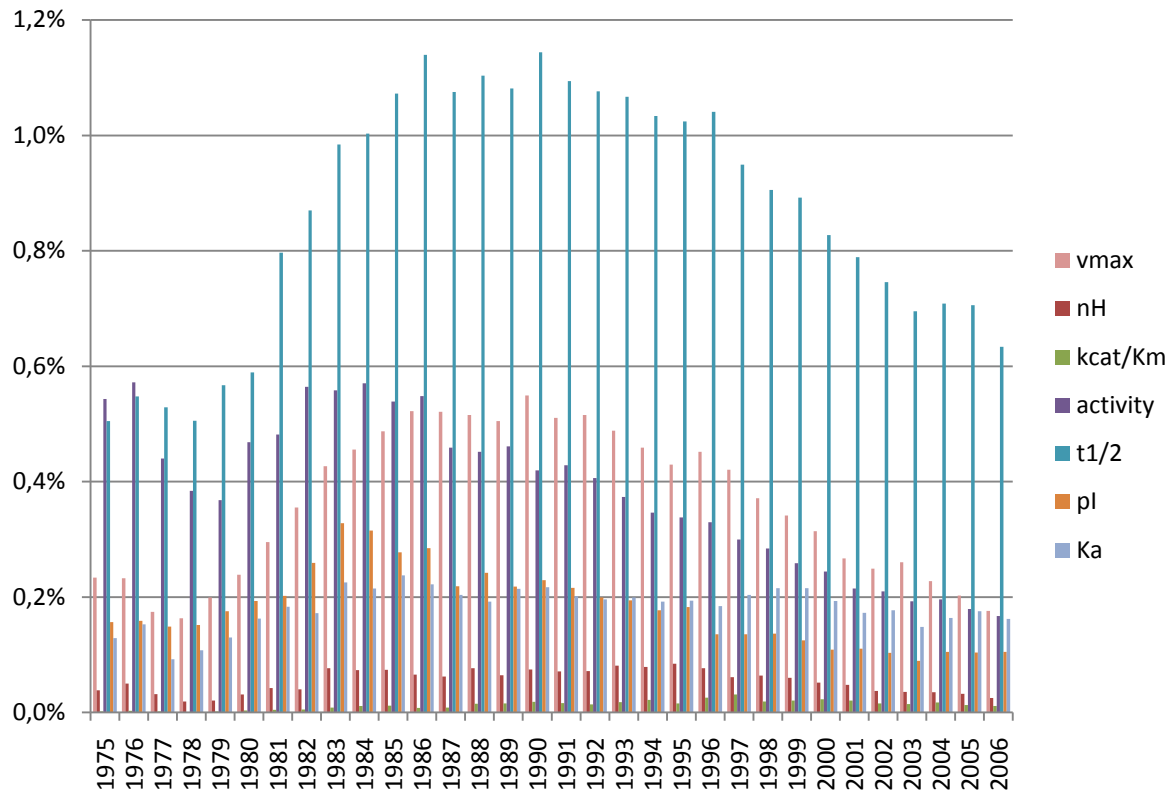

Abbildung 35: Menge an Abstracts bezogen auf den beinhalteten kinetischen Ausdruck und die Menge an veröffentlichten Abstract im Erscheinungsjahr Teil 2 (siehe Anhang Punkt 7.1.6).

3.6. Qualitative Analyse der Ergebnisse

Für die qualitative Analyse wurden 1.002 Einträge aus verschiedenen zufällig ausgewählten Abstracts validiert. Hierbei wurde für jede Kategorie vermerkt, ob der Wert falsch, richtig, im Abstract enthalten, aber nicht gefunden wurde oder gar nicht enthalten ist.

In Abbildung 36 ist der Recall (blaue Balken) für die verschiedenen Kategorien graphisch aufgetragen.

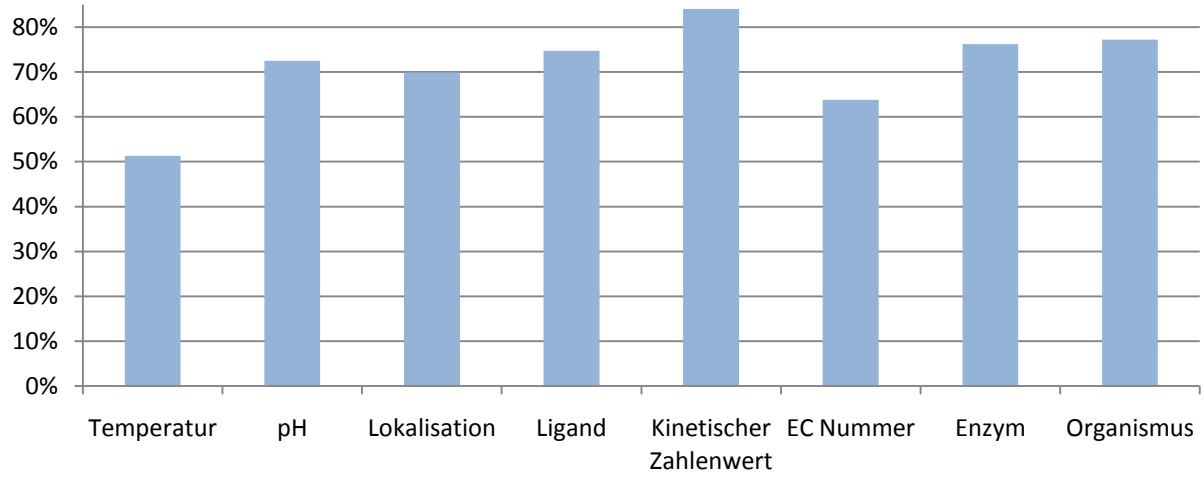

Abbildung 36: Graphische Darstellung des Recalls unterteilt nach Kategorien (siehe Anhang Punkt 7.1.7).

Der größte Recall wird für Liganden (75%) und Organismen (77%) erzielt. Für Lokalisationen, Zahlenwerte, pH-Werte und Enzymnamen wird ein Recall von 70%, 84%, 72% und 76% erreicht. Der geringste Recall wird für EC Nummern (63%) und Temperaturen (51%) ermittelt.

Die für die verschiedenen Kategorien erhaltene Precision ist in Abbildung 37 dargestellt.

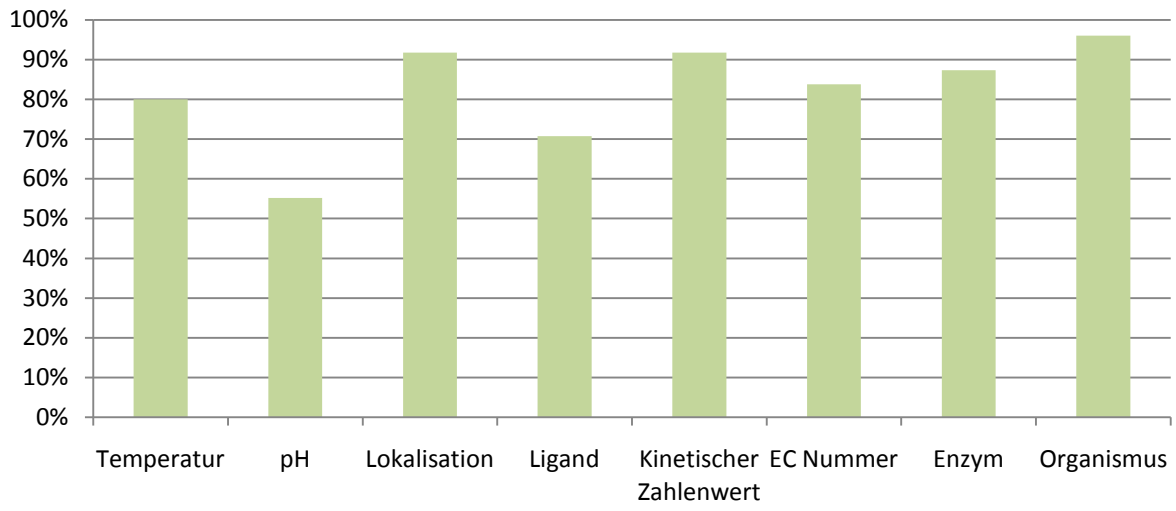

Abbildung 37: Graphische Darstellung der Precision-Werte unterteilt nach Kategorien (siehe Anhang Punkt 7.1.7).

Die größte Precision zeigt sich bei Organismen (96%) und Lokalisationen (92%). Für Zahlenwerte liegt die Precision bei 92%. 87%, 84% und 80% wurden für Enzymnamen, EC Nummern bzw. Temperaturen erreicht. Die Precision für Liganden liegt bei 71%. Die geringste Precision wurde für pH-Werte mit 55% ermittelt.

Abbildung 38 gibt einen Überblick über die Genauigkeit pro ermittelten Eintrag im Zusammenhang mit dem enthaltenen kinetischen Ausdruck.

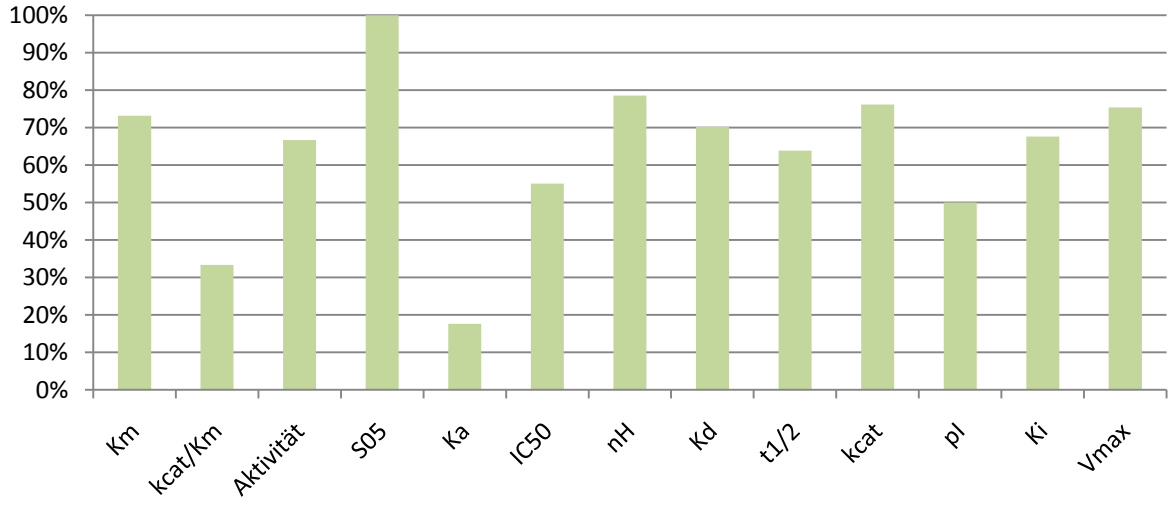

Abbildung 38: Genauigkeit der gesamten Einträge unterschieden nach den beinhalteten kinetischen Ausdrücken. 100% entspricht 149 (K_M), 6 (k_{cat}/K_M), 69 (spezifische Aktivität), 4 ($S_{0.5}$), 17 (K_a), 198 (IC_{50}), 14 (n_H), 114 (K_d), 202 ($t_{1/2}$), 21 (k_{cat}), 30 (pI), 102 (K_i) und 73 (V_{max}) (siehe Anhang Punkt 7.1.7).

Der Großteil der anderen Kategorien weist eine Genauigkeit von 50% bis 70% auf.

3.7. Vergleich mit anderen Datenbänken

Die folgende Tabelle zeigt einen inhaltlichen Vergleich mit den Datenbänken KMedDB [30] und BRENDA [6], die ebenfalls eine Sammlung an kinetischen Informationen beinhalten.

Tabelle 11: Inhaltlicher Vergleich mit KMedDB und BRENDA

	KID		KMedDB	BRENDA
Algorithmus	Basiert auf Linguistik; lexikagestützt		Machine Learning [30]	Manuell; seit 1987 [6]
Datenquelle	Abstracts		Abstracts [30]	Abstracts und Artikel [6]
Recall	51% bis zu 84%		40% bis zu 63% (Training) [30] 31% (unbekannte Daten) [30]	---
Precision	55% bis zu 96%		100% (Training) [30] 50% (unbekannte Daten) [30]	----
Kinetischer Ausdruck	**Anzahl Abstracts**	**Anzahl Einträge**	**Anzahl Abstracts**	**Anzahl Einträge**
$t_{1/2}$	57.658	91.429	49.608 [35]	---
IC_{50}	54.938	105.24	28.709 [35]	---
K_M	45.896	74.253	39.719 [35]	76.966 [26]
K_d	40.244	65.401	39.448 [35]	---
V_{max}	23.985	36.862	29.851 [35]	---
K_i	22.805	42.344	20.683 [35]	15.564 [26]
k_{cat}	7.325	10.405	9.528 [35]	23.798 [26]
Unterschiedliche kinetische Kategorien	n_H, k_{cat}/K_M, K_a, $S_{0.5}$, pI, spezifische Aktivität		n_H, k_{cat}/K_M, K_a, K_{eq} [35]	spezifische Aktivität, pI [26]
Summe	260.316	514.394	199.701 [35]	145.791 [26]

Im Vergleich mit KMedDB und BRENDA zeigen sich Unterschiede im Umfang. KMedDB wurde mittels eines Textminingalgorithmus unter Verwendung von Machine Learning aus Abstracts erstellt wohingegen BRENDA seit 1987 manuell aus Abstracts und Artikeln annotiert wird. Für KMedDB wird ein Recall von 40% bis 63% und eine Precision von 100% für den Trainingsdatensatz angegeben. Für einen unbekannten Datensatz werden ein Recall von 31% und eine Precision von 50% ermittelt.

Durch Vergleich der Menge an enthaltenen Abstracts wird ersichtlich, dass in KID mehr Abstracts enthalten sind als in KMedDB. Im Vergleich mit BRENDA zeigt sich, dass für K_M eine ähnliche Anzahl an Informationen extrahiert werden konnte. Für K_i und k_{cat} unterscheiden sich die Mengen.

4. Diskussion

4.1. Recall und Precision

Insgesamt konnten Informationen zu 509.153 verschiedenen kinetischen Ausdrücken mit einer Precision von 55% bis zu 96% und einem Recall von 51% bis zu 84% extrahiert werden (vergleiche Abschnitt 3.6).

Der Recall macht eine Aussage darüber, wie viele der im Gesamtbestand enthaltenen Informationen extrahiert werden [11, 17]. Die Precision ist ein Maß dafür, wie viele dieser extrahierten Informationen relevant bzw. richtig sind.

Um eine Aussage über diese beiden Werte treffen zu können, wurden 1002 Einträge zufällig ausgewählt und manuell für alle Kategorien bewertet. Hierbei wurde vermerkt, ob eine inhaltlich mit einem kinetischen Ausdruck verbundene Information aus der jeweiligen Kategorie im Abstract enthalten ist oder nicht. Wurde eine Entität extrahiert, wurde vermerkt, ob sie inhaltlich richtig oder falsch ist. Hierbei haben sich Recall-Werte [6, 38] von ca. 51% bis 84% für ergeben (siehe Abbildung 36).

Auch die ermittelten Werte für die Precision sind mit 55% bis 96% (siehe Abbildung 37) relativ hoch (vergleiche 4.2), was darauf schließen lässt, dass der Algorithmus mit hoher Genauigkeit bei der Extraktion und Verbindung der Werte vorgeht. Dies ist auf die strengen Regeln bei der inhaltlichen Verbindung zurückzuführen. So wird durch die Verwendung des Kommas als Stopsignal gewährleistet, dass in Fällen wie z.B. „…Km is 3 M for ATP, 4 M for GTP…“ „ATP“ trotz einer direkten Nachbarschaft nicht mit „4 M“ verknüpft wird. Durch das Ignorieren des Kommas, wenn dieses von bestimmten Schlüsselworten gefolgt wird, z.B. „which“, wird wiederum das inhaltliche Verbinden von Werten im nächsten Nebensatz ermöglicht, wodurch der Recall erhöht, aber die Precision nicht merklich verringert wird. Auch wird bei dem Vorkommen eines Kommas zwischen Markierung einer Satzebene und Markierung einer Auflistung unterschieden und entsprechend verfahren. Im Fall der

Auflistung werden alle ihre Mitglieder zu einem Wert oder zu unterschiedlichen Mitgliedern einer anderen Auflistung zugeordnet.

Es wurde die Genauigkeit aller Kategorien pro Eintrag, bzw. ob in einem Eintrag eine falsche Information enthalten ist, untersucht, wodurch sich eine Aussage über die „Precision“ der Kombination extrahierter Werte machen lässt (siehe Abbildung 38). Für die kinetischen Kategorien, zu denen mit die meisten Informationen extrahiert wurden ($t_{1/2}$, K_d und K_M), wurde eine Genauigkeit von 64%, 70% bzw. 73% für 202, 114 bzw. 149 Einträge ermittelt, was darauf hinweist, dass ein Großteil der Kombinationen extrahierter Werte eine richtige Aussage über die kinetische Information im Abstract macht. Für Kategorien, die in geringen Mengen in der Validierung auftraten, wurden keine statistisch relevanten Werte ermittelt.

4.2. Evaluierung des Algorithmus

Anhand der oben aufgeführten Untersuchungen zu Recall und Precision ist erkennbar, dass mittels des vorgestellten Algorithmus viele Ergebnisse mit guter Qualität aus einem unbekannten Datensatz extrahiert werden können. Durch die Verwendung von Lexika unter Berücksichtigung grammatikalischer Regeln wurde eine pragmatische Annäherung umgesetzt (siehe 1.1 bis 1.4).

Im Gegensatz zu diesem Algorithmus werden in der Forschung oft Machine Learning bei der Entwicklung von Textminingalgorithmen zur Hilfe gezogen [30, 34]. Ergebnisse dieses Verfahrens sind z.B. in den Datenbanken KMedDB [31] und Kinetikon [32] zugänglich. Ein bekannter Nachteil besteht jedoch darin, dass für den Trainingsdatensatz gute Ergebnisse erzielt werden (bis zu 100% Precision, Recall ca. 40% bis 63%), bei unbekannten Datensätzen können Recall und Precision jedoch sinken (Precision 50%, Recall ca. 31%) [30]. Bei der Verwendung von Lexika kommt ein ähnlicher Phänomen zum Tragen, da die manuell ergänzten Einträge aus dem verwendeten Datensatz stammen. Da das Auffinden neuer Entitäten, vor allem solcher mit häufigem Auftreten (z.B. ATP), mit wachsenden Lexika unwahrscheinlicher werden sollte, ist davon auszugehen, dass der Effekt weniger ausgeprägt ist.

Der Vergleich mit Programmen [37, 38] und Datenbanken [6, 36], deren Ergebnisse aus der automatischen Suche über Abstracts stammen, zeigt, dass die Werte für Recall und Precision in ähnlichen Bereichen liegen [6, 36] oder höher sind. So erreicht BioRAT [37] einen Recall

von ca. 20% und eine Precision von ca. 55%. Für FRENDA und AMENDA wurden jeweils ein Recall von 72% und 11,7% sowie eine Precision von 64,8% und ~76% angegeben.

Durch die Rechenzeit von 0,004 Sekunden pro Abstract ist es mit dem vorgestellten Algorithmus möglich, auch große Datenmengen in einer annehmbaren Zeit auf kinetische Daten zu untersuchen. Die Rechenzeit resultiert vor allem aus der Verwendung des Lexikon, das durch seine Hash-basierte Struktur eine konstante Laufzeit nach der O-Notation O(n) [47] bei der Identifizierung garantiert (siehe Abbildung 8 und Abbildung 9). Dadurch, dass das Identifizieren der Zahlen nur in Abstracts erfolgt, die auch einen kinetischen Ausdruck beinhalten, wird die Rechenzeit weiter optimiert. Obwohl die Suche mit einem regulären Ausdruck zeitaufwendiger ist, bietet es sich an dieser Stelle an, da Zahlen größtenteils in festgelegten Schreibweisen verwendet werden. Eine lexikongestützte Identifikation wäre in diesem Fall nicht sinnvoll, da die Menge an möglichen Zeichenfolgen unendlich groß ist. Bei der inhaltlichen Verbindung der extrahierten Entitäten werden bei der direkten Verbindung im Satz mit dem kinetischen Ausdruck nur die Entitäten betrachtet, die durch verbindende Ausdrücke miteinander in Verbindung stehen (siehe Abbildung 10 bis Abbildung 14). Somit wird gewährleistet, dass außer im Fall einer Auflistung, jede Entität einmal untersucht wird. Bei der indirekten Verbindung, wird anschließend nach noch nicht gefundenen Kategorien gesucht (siehe Abbildung 15), so dass hier die Anzahl der zu überprüfenden Kategorien bereits durch die direkte Verbindung verringert wurde. Da die inhaltliche Verbindung zwar rechenintensiver ist, aber nur in den Abstracts durchgeführt wird, die auch einen kinetischen Ausdruck beinhalten (260.316 Abstracts), geschieht dies nur in 1,54% der Gesamtmenge der zu untersuchender Abstracts. Im Gegensatz hierzu muss die Identifikation der Entitäten jedoch in allen 16.953.021 Abstracts durchgeführt werden.

Für BioRAT [37] ist eine durchschnittliche Rechenzeit von 3 – 5 Sekunden pro Abstract bei Verwendung vergleichbarer Hardware (0,5 GB Ram, 1,7 GHz, Single Core von 2004) angegeben. Durch Hochrechnen der angegebenen Rechenzeit auf den verwendeten Datensatz ergibt sich eine Gesamtrechenzeit von etwa 600 – 1.000 Tagen, was nahezu drei Größenordnungen höher ist als die Rechenzeit des hier vorgestellten Algorithmus und damit eine Bearbeitung umfassender Datensätze in ansprechender Zeit unmöglich macht.

Durch die Verwendung definierter Suchbegriffe in den Lexika besteht die Möglichkeit, dass Begriffe, die nicht in dem Lexikon enthalten sind, nicht aus dem Text extrahiert werden können. Ein möglichst komplettes und umfassendes Lexikon ist daher wünschenswert, so dass das Einpflegen von Begriffen auch in Zukunft erforderlich sein wird. Auch stellen

Rechtschreibfehler und verschiedene Schreibweisen von Begriffen, wie z.B. „glucose-6-phosphate“ und „glucose 6 phosphate“, im Abstract ein Problem bei der Wertextraktion dar. Zudem ist im Fall einer Synonymie eine eindeutige Zuordnung des Begriffes nur schwer möglich. Aus diesem Grund wurden diese Begriffe, wie „IMP“ (Abkürzung für den Liganden „Inositol monophosphate“ aber auch für das Enzym „Inositol-phosphate phosphatase; EC 3.1.3.25“) in Ausschlusslisten manuell gesammelt. Hier sollte in späteren Versionen eine eventuelle Untersuchung der umgebenden Wörter, z.B. „….the ligand IMP“ oder „…the enzyme IMP“, eine genauere Zuordnung des Begriffes zu einer Kategorie ermöglichen. Da diese Fälle jedoch relativ selten im Vergleich zur extrahierten Gesamtmenge vorkommen, verzichtet der Algorithmus auf eine solche Untersuchung der Umgebung. Zusätzlich erschweren Rechtschreibfehler die Identifikation von Entitäten. In solchen Fällen wäre eine Erkennung von Rechtschreibfehlern z.B. aufgrund von Wortähnlichkeiten in späteren Versionen sinnvoll, so dass auch falsch geschriebene Entitäten identifiziert werden können.

Da die Lexika ursprünglich nur aus Nomen, wie z.B. „kidney“, bestanden, wurden zusätzlich Einträge in Adjektivform manuell hinzugefügt. Auf diese Weise konnte die Menge an falsch Negativen, d.h. Fälle, in denen eine Information im Text enthalten ist, aber nicht erkannt wurde, deutlich gesenkt werden. So ist beispielsweise in „….The Km of the renal enzyme was 720 ng/ml, of the enzyme prepared from…“ [48] die Information über das Gewebe als Adjektiv „renal“ enthalten. Durch Aufnahme dieses Beispiels in das Lexikon für Gewebe erfolgte in 3.741 Fällen eine Verknüpfung.

4.3. Präsentation der Ergebnisse

Durch die Präsentation der Ergebnisse im Onlineinterface werden die Ergebnisse in einer benutzerfreundlichen Art zugänglich gemacht. Die Suchfunktion ermöglicht die gezielte Entnahme von Informationen z.B. zu einem Enzym oder Organismus aus der Datenbank. Durch das Einfärben der Ergebnisse im Originaltext wird dem Nutzer das Erfassen der inhaltlichen Zusammenhänge zwischen den einzelnen Entitäten erleichtert und somit deren Kontrolle beschleunigt.

Die Validierung gibt die Möglichkeit, Fehler in folgenden Versionen zu korrigieren und fehlende Begriffe in die Lexika aufzunehmen.

4.4. Arten der inhaltlichen Verbindung

Die statistische Untersuchung über die Häufigkeit der verschiedenen inhaltlichen Verbindungen wurde für IC_{50} und K_M beispielhaft durchgeführt, da zu diesen beiden kinetischen Kategorien die meisten Ergebnisse extrahiert wurden. Hierbei zeigt sich, dass bei den Kategorien unterschiedliche Arten der inhaltlichen Verbindung bevorzugt werden (Abbildung 24 und Abbildung 25). So wird die direkte Verbindung überwiegend für die Verknüpfung mit Liganden (30% und 37%) und Zahlenwerten (59% und 57%) genutzt. Häufig befinden sich Zahlenwerte auch in Form von Auflistungen im Satz (32% und 26%). Auch die indirekte Zuweisung auf der Satzebene (34% und 26%) trifft in dem Fall des Liganden relativ oft zu. Aus diesen Werten lässt sich schließen, dass Entitäten dieser beiden Kategorien sich häufig in einem Satz mit dem kinetischen Ausdruck befinden.

Im Gegensatz hierzu ist die Suche über den gesamten Abstract und über den Titel besonders erfolgreich bei der Zuweisung von Enzyminformationen, Organismen und Lokalisationen. In diesen Fällen werden bis zu 83% der Entitäten indirekt zugewiesen. Zusammen mit der Precision (siehe Abbildung 37) lässt sich hieraus schlussfolgern, dass die Nennung dieser Informationen meist nur einmal im Abstract oder Titel erfolgt und sich die darin enthaltenen Informationen auf den kinetischen Ausdruck beziehen.

Die automatische Ergänzung findet vor allem bei EC Nummern anhand eines extrahierten Enzymnamens mit ca. 72% bis 92% statt. Die Zahl der automatisch annotierten Enzymnamen anhand einer extrahierten EC Nummer ist verschwindend gering. Beides hat seine Ursache darin, dass Enzymnamen häufiger als EC Nummern im Text auffindbar sind. In Fällen, in denen eine inhaltliche Verbindung zu einem Rezeptor gegeben ist, kann oftmals keine Zuweisung zu einer EC Nummer erfolgen, da diese oftmals keine EC Nummer besitzen.

Eine Schwierigkeit ergibt sich jedoch bei der indirekten Verbindung, wenn sich in einem Satz verschiedene Synonyme zu einem Begriff einer Kategorie befinden. Die indirekte Verbindung wird nicht vorgenommen, obwohl inhaltlich nur ein Objekt beschrieben wird. Abhilfe könnte eine Synonymerkennung schaffen, die sich jedoch negativ auf die Laufzeit des Programms auswirken würde.

4.5. Extrahierte Ergebnisse

Bei der Analyse der extrahierten Informationen zeigt sich, dass die meisten Informationen zu IC_{50} (105.240 Einträge), $t_{1/2}$ (91.430 Einträge), K_M (74.253 Einträge) und K_d (65.401 Einträge) gefunden werden (vergleiche Abbildung 18 und Abbildung 19). Für k_{cat}/K_M werden mit 1.469 die wenigsten Informationen gesammelt.

Der Vergleich mit der in KMedDB [31] enthaltenen Menge an Abstracts mit kinetischem Inhalt zeigt, dass durch das hier vorgestellte Programm für die meisten kinetischen Ausdrücke eine ähnliche große Menge an Literaturstellen gefunden wurde. So sind zu K_M in der KMedDB ca. 40.000 Abstracts aus 15 Millionen Abstracts [35] und in KID ca. 45.000 Abstracts aus ca. 17 Millionen Abstract enthalten. Im Fall von IC_{50} wurden in KID ca. 45.000 Abstracts gefunden, wohingegen in KMedDB ca. 29.000 Abstracts zu diesem Wert vermerkt sind (siehe Tabelle 11).

Vergleicht man die in dieser Arbeit erstellte KID-Datenbank mit der BRENDA [26], so zeigt sich, dass für die Michaelis-Menten Konstante eine annähernd gleich große Menge an Informationen extrahiert werden konnte (siehe Tabelle 11). Für K_i war es sogar möglich mehr Informationen (42.344 Einträge) als in BRENDA enthalten (15.564 Einträge) zu ermitteln, was sich vor allem auf die Tatsache zurückführen lässt, dass Informationen zu diesem Wert erst seit zwei Jahren in BRENDA [26] annotiert werden.

Die nachfolgende Tabelle zeigt einen Vergleich der in KID und BRENDA enthaltenen Artikel. Für K_M, k_{cat} und K_i war es möglich eine Schnittmenge zwischen beiden Datenbanken zu ermitteln, dennoch sind in KID eine beträchtliche Menge an Artikel enthalten, die nicht in BRENDA annotiert sind. Textmining bildet somit auch eine hilfreiche Ergänzung zur manuellen Annotation von kinetischen Werten.

Tabelle 12: Vergleich der Menge an Artikeln mit BRENDA

	KID	BRENDA*	Schnittmenge
K_M	45.896	13.541	5.115
K_i	22.805	2.600	793
k_{cat}	7.325	3.064	830

* BRENDA von 2005

Auch für die verschiedenen Suchkategorien (siehe Abbildung 18 und Abbildung 19) konnten große Mengen an Informationen zugewiesen werden (z.B. Verknüpfungen zu 338.706

Liganden insgesamt). Auffällig ist hier, dass die Menge an Enzyminformation (Rezeptorname oder Enzymname und EC Nummer) bei $t_{1/2}$ relativ selten vertreten ist, was sich dadurch erklären lässt, dass dieser Wert häufig eine physiologische oder biologische Halbwertszeit darstellt, die sich nicht auf ein Enzym bezieht (siehe 1.7.7).

Zudem fällt auf, dass Informationen zu pH und der Temperatur nur in vergleichsweise geringen Mengen extrahiert werden konnten. Dies kann zum einen darauf zurückzuführen sein, dass die verwendeten Messbedingungen eher im Paper als im Abstract erwähnt werden, zum anderen können diese Informationen auch implizit z.B. im verwendeten Medium enthalten sein. In Fällen, in denen beispielsweise ein Enzym im Magensaft gemessen wird, kann der Leser von dem Magensaft auf einen pH von 1 schließen, wohingegen bei Messungen im Blut, auf einem physiologischen pH-Wert von pH 7 zu schließen ist.

Da die biologische Information in der Kombination von Entitäten verschiedener Kategorien liegt, wurden die Mengen der Kombinationen, die eine relevante Aussage enthalten untersucht. Hierbei (siehe Abbildung 20 bis Abbildung 23) zeigt sich, dass mit der Anzahl der verbunden Kategorien der prozentuale Anteil der Kombination sinkt. Insgesamt konnten weniger umfangreiche Kombination (beispielsweise ca. 60% für $S_{0.5}$ mit Enzym und Ligand) in mehr als der Hälfte der Fälle zugewiesen werden.

4.6. Verteilung von extrahierten kinetischen Zahlenwerten

Eine genauere Untersuchung der Zahlenwerte zeigt, dass für jeden kinetischen Ausdruck ein Präferenzbereich ermittelt werden konnte. Für K_d wurde so z.B. ein Maximum bei 10^{-6} bis 10^{-5} mM ersichtlich. Angaben zu Bindungseigenschaften unterscheiden sich somit eindeutig von Informationen zu K_M, K_i, $S_{0.5}$ und IC_{50}, deren Zahlenwerte großteilig in dem Bereich von 10^{-3} bis 10^{-1} mM liegen. Auch für Werte anderer Kategorien lassen sich Präferenzbereiche angeben.

4.7. Verteilung von extrahierten EC Nummern

Bei der genaueren Betrachtung der ermittelten EC Nummern ist auffällig, dass die meisten Informationen zu Hydrolasen (EC Klasse 3) gehören (siehe Tabelle 9). Da jedoch bereits im Lexikon die meisten Einträge zu dieser Klasse gehören, lässt sich nicht auf eine Sonderstellung dieser Klasse in den Ergebnissen schließen.

4.8. Verteilung von extrahierten Organismen

Bei der Verteilung der extrahierten Organismen zeigt sich ein Schwerpunkt bei Tieren, bzw. Säugetieren (siehe Abbildung 31). Da die häufigsten Organismen Menschen, Maus und Ratte sind (siehe Tabelle 10), kann dies in diesem Zusammenhang darauf hinweisen, dass ein Großteil der Abstracts einen medizinisch-pharmakologischen Hintergrund besitzt. Dies zeigt, wie wichtig die Datenbank für die Medizin und Pharmakologie ist.

4.9. Verteilung der Publikationsdaten

Die Notwendigkeit eines Textminingprogramms zur Extraktion kinetischer Informationen aus der Literatur wird durch die Analyse der Publikationsdaten von Abstracts mit kinetischem Inhalt offensichtlich (siehe Abbildung 32 bis Abbildung 35). In diesen beiden Abbildungen zeigt sich, dass in den letzten Jahren die Anzahl an Publikationen mit kinetischem Inhalt gestiegen ist und in Zukunft wahrscheinlich noch weiter ansteigen wird. Es wird deutlich, dass die Menge von jährlich ca. 19.000 publizierten Informationen die Möglichkeiten einer rein manuellen Suche überschreitet.

Bei der Betrachtung der zeitlichen Verteilung der kinetischen Informationen ist auffällig, dass die Parameter K_M, K_i, K_d, V_{max} in den 1980er bis 1990er Jahren eine Hochzeit hatten und seitdem etwas abgefallen sind. Eine mögliche Erklärung hierfür könnte die Entwicklung neuer Methoden wie der NMR (1985, [42]) oder Verfahren zur Analyse und Amplifikation erwünschter Genprodukte wie der PCR (1984, [3, 44]) sein, die die Messung von kinetischen Werten erleichterten [44]. Im Zusammenhang mit der röntgenkristallographischen Aufklärung von Proteinen, die seit Anfang der 1990er Jahre stetig zunimmt, werden häufig Bindungskonstanten von Liganden bestimmt, was ebenfalls zu der beobachteten Zunahme beitragen könnte [40]. Ein weiterer Grund für die Steigerung könnte in der Zunahme von Bindungsstudien zur Signaltransduktion in Eukaryoten liegen, die seit den 1980ern Jahren an Bedeutung gewannen, z.B. durch die Aufklärung von Mutationen in Entwicklungsgenen von *Drosophila melanogaster* [39].

Für $t_{1/2}$ lässt sich ein Anstieg der Publikationsdaten von 1980 bis 2000 feststellen. Diese Phase wird von einer weitgehend gleichbleibenden Menge an Publikationen gefolgt, was auf eine gleichbleibende Menge an pharmakologischen Studien hinweist.

Das Maximum an Publikationen zu IC_{50} liegt im Gegensatz zu den oben genannten Parametern in den 1990ern Jahren. Dieser Begriff findet jedoch weniger in der Molekularbiologie, sondern häufiger in der Pharmakologie Verwendung [29], was die Verschiebung erklären könnte.

Eine weiterführende Analyse der Publikationsmengen bezogen auf das Erscheinungsjahr könnte genauere Einblicke in die Hintergründe liefern, setzt aber eine manuelle Auswertung einer großen Zahl von Artikeln voraus und liegt damit nicht im Rahmen dieser Arbeit.

5. Schlussfolgerung

Ziel dieser Arbeit war die Entwicklung eines Programms zur automatischen Extraktion kinetischer Werte. Mit dem hier vorgestellten Algorithmus ist es möglich in kurzer Zeit große Mengen an kinetischen Informationen mit hoher Genauigkeit zu extrahieren, womit er eine gute Alternative zu anderen Algorithmen, z.B. solche die auf Maschine Learning basieren, darstellt. Die erhaltenen Ergebnisse wurden hinsichtlich ihres Zahlenwertes, ihres Organismus, ihrer EC Klasse und ihres Publikationsdatums untersucht und eine Analyse hinsichtlich der Vollständigkeit und Korrektheit durchgeführt.

Durch Weiterentwicklungen im Bereich des Textminings wird es in Zukunft möglich sein, umfangreiche Informationen aus Fließtexten in vielerlei Bereichen zu gewinnen und auszuwerten.

6. Literatur

1. „Enzymkinetik“ von H. Bisswanger, WILEY-VCH Verlag GmbH, Weinheim, Deutschland, 2000, ISBN 3-527-30096-1
2. „Prinzipien der Biochemie“ von Lehninger, Nelson und Cox; Spektrum Verlag, Berlin, Deutschland, ISBN 3-8274-0325-1, 2. Auflage
3. „Biochemie“ von L. Struyer, Spektrum Verlag, Heidelberg, Deutschland, 1996, ISBN 3-86025-346-8, 4. Auflage
4. „The possible effects of the aggregation of molecules of hemoglobin on its dissociation curves“ von A. V. Hill, The Journal of Physiology 40, 4-7,1910
5. „Die Kinetik der Interverinwirkung” von L. Michaelis und M. L. Menten, Biochemische Zeitschrift 49, 1913, 333-369
6. „BRENDA, AMENDA and FRENDA: the enzyme information system in 2007“ von J. Barthelmes, C. Ebeling, A. Chang, I. Schomburg und D. Schomburg, Nucleic Acids Research, 2007, Vol. 35, Database issue D511–D514
7. “BRENDA, enzyme data and metabolic information“ von I. Schomburg, A. Chang und D. Schomburg; Nuclein Acids Research, 2002, Vol. 30, Oxford University Press
8. “BRENDA: a resource for enzyme data and metabolic information” von I. Schomburg, A. Chang, O. Hofmann, C. Ebeling, F. Ehrentreich und D. Schomburg, Trends in Biochemical Science; Vol. 27 , Januar 2002
9. http://www.bio.ifi.lmu.de/forschung/textmining
10. „Automatische Textanalyse Systeme und Methoden zur Annotation und Analyse natursprachlicher Texte“ von A. Mehler und H. Lobinger, VS Verlag für Sozialwissenschaften, Wiesbaden, Deutschland, 2004, ISBN 3-531-14181-3
11. „Grundlagen der Computerlinguistik Mensch-Maschine-Kommunikation in natursprachlicher Sprache“ von R. Hausser, Springer Verlag, Berlin Heidelberg, Deutschland, 2000, ISBN 3-540-67187-0

12. „Text Mining for Biology and Biomedicine“ von S. Ananiadou und J. McNaught, Artech House, London und Boston, 2006, ISBN 1-58053-983-x
13. „Untagling text data mining“ von M. Hearst, In proceedings of ACL, 1999 Pages 3-10
14. http://www.innovations-report.de/html/berichte/biowissenschaften_chemie/bericht-17197.html
15. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed
16. „English G Grammatik” von E. Fleischhack, H. Schwarz und F. Vettel, Cornelsen Verlag, Berlin, Deutschland, ISBN 3-464-00372-8
17. „Text Mining: Wissensrohstoff Text“ von G. Heyer, U. Quasthoff und T. Wittig, W3L-Verlag, Bochum, Herdecke, Deutschland, ISBN 3-937137-30-0
18. „Analytische Philosophie I“ von W. K. Essler, Kröner Verlag, Stuttgart, Deutschland, ISBN 3-520-44001-6
19. „Wörterbuch der Philosophie“ von G. Klaus und M. Buhr, Rowohlt-Verlag, Reinbek beim Hamburg, Deutschland, ISBN 3-499-16157-5
20. „Neurowissenschaften“ von E. R. Kandel, J. H. Schwartz und T. M. Jessell; Sprektrum-Verlag, Berlin, Deutschland, ISBN 3-86025-391-3
21. „Interaktionen zwischen dem Peptidhormon Relaxin und dem humanen Glukokortikoidrezeptor“ von Michael Peter Greinwald, Dissertation, vorgelegt der Medizinischen Fakultät der Charité – Universitätsmedizin Berlin, 2006, http://edoc.hu-berlin.de/dissertationen/greinwald-michael-peter-2006-05-15/HTML/
22. http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmeddata.htm
23. http://www.expasy.org/cgi-bin/lists?tisslist.txt
24. http://www.trolltech.com/products/qt
25. http://www.joomla.org/
26. http://www.brenda-enzymes.info/
27. http://www-ra.informatik.uni-tuebingen.de/lehre/ss02/pro_wirkstoffdesign_ausarbeitung/nina_fischer.pdf
28. http://amor.rz.hu-berlin.de/~h2816i3x/GK_Semantik_06_Bedeutungsbeziehungen.pdf
29. „Antiprotozoal and cytotoxic naphthalene derivatives from Diospyros assimilis.” von S. Ganapaty, P. S. Thomas, G. Karagianis, PG. Waterman und R. Brun, Phytochemistry. Vol. 67, 2006, 1950-6
30. „Text Mining for Sytem Biology Using Statistical Learning Methods“ von S. Schmeier, J. Hakenberg, A. Kowald, E. Klipp und U. Leser; http://km.aifb.uni-karlsruhe.de/ws/LLWA/akkd/7.pdf

31. http://sysbio.molgen.mpg.de/KMedDB
32. http://kinetikon.molgen.mpg.de/
33. „Automated recognition and extraction of entities related to enzyme kinetics from text", Master´s Thesis von Sebastian Schmeirer, Berlin, 2005, http://www2.informatik.hu-berlin.de/~hakenber/theses/schmeier_the051031.pdf
34. „Finding kinetic parameters using text mining." von J. Hakenberg, S. Schmeier, A. Kowald, E. Klipp und U. Leser, OMICS, Vol 8, 2004, 131-52
35. http://www.molgen.mpg.de/~schmeier/data/KMedDB-dayOfScience.pdf
36. http://www.molgen.mpg.de/~schmeier/data/BIOTEXT-Poster.pdf
37. „BioRAT: extracting biological information from full-length papers" von David P. A. Corney, Bernard F. Buxton, William B. Langdon und David T. Jones, Bioinformatics, Vol. 20, 2004, 3206–3213
38. „A survey of current work in biomedical text mining" von A. M. Cohen und W. R. Hersh, Briefings in bioinformatics, Vol. 6, 2005, 57-71
39. „Mutations affecting segment number and polarity in Drosophila." von C. Nüsslein-Volhard und E. Wieschaus, Nature, Vol. 287, 1980795-801
40. http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=explMethod-xray&seqid=100
41. http://www.ncgc.nih.gov/guidance/section3.html
42. „Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry " von M. P. Williamson, T. F. Havel und K. Wuthrich, J. Mol. Biol., Vol. 182, 1985, 295-315
43. „Pharmakologie und Toxikologie" von W. Forth, D. Henschler, W. Rummel und K. Starke, Spektrum Verlag, Heidelberg, Deutschland, ISBN 3-8274-0088-0, 7. Auflage
44. „A History of Molecular Biology" von Michel Morange, Harverd University Press, London, England, 2000, ISBN 0-674-00169-9, 1. Auflage
45. „Relationship between in vivo degradative rates and isoelectric points of proteins" von J. Fred Dice und Alfred L. Goldberg, Proc. Nat. Acad. Sci., Vol. 72, 1975, 3893-3897
46. http://www.w3.org/TR/soap/
47. „Algorithmen" von R.Sedgewick, Pearson Studium, München, Deutschland, 2002, ISBN 3-8273-7032-9
48. „Kinetic differences between uterine and renal renins in the dog" von Potter et al., Am J Physiol Endocrinol Metab., 1977, Vol. 233

7. Anhang

7.1. Daten

7.1.1. Extrahierte Mengen aller Kategorien

	Gesamtanzahl kinetischer Ausdruck	Ligand	Enzym	Organismus	Lokali= sation	EC	Zahlen= wert	pH	Temp.
IC_{50}	105.240	72.493	55.321	69.931	63.526	12.114	72.714	402	444
$t_{1/2}$	91.430	63.047	22.252	52.077	53.024	6.004	45.371	1.101	2.961
K_M	74.253	54.196	51.277	53.326	39.437	19.461	33.879	2.642	1.997
K_d	65.401	41.230	35.087	41.958	36.998	5.460	35.973	1.065	1.947
Ki	42.344	29.526	29.050	26.747	21.024	8.367	27.210	633	427
Spez. Aktivität	37.001	21.739	23.558	27.011	22.373	9.972	3.402	497	745
V_{max}	36.862	22.546	21.241	25.796	21.899	6.500	4.364	678	838
K_a	18.903	11.978	7.184	10.786	9.559	1.289	3.246	342	636
pI	17.481	8.969	8.213	11.854	9.425	2.997	10.158	1.855	246
k_{cat}	10.405	6.289	7.806	6.374	2.848	3.033	1.952	622	568
n_H	5.465	3.458	3.333	3.779	2.929	746	3.978	127	86
$S_{0.5}$	2.899	2.264	2.143	1.969	1.714	706	1.802	114	39
k_{cat}/K_M	1.469	971	891	971	697	352	75	34	11
Summe	509.153	338.706	267.356	332.579	285.453	77.001	244.124	10.112	10.945

7.1.2. Kombinationen extrahierter Werte

	K_M	K_i	k_{cat}	$S_{0.5}$	IC_{50}	n_H	K_d	V_{max}
Gesamtanzahl des kinetischen Ausdrucks	74.253	42.344	10.405	2.899	105.240	5.465	65.401	36.862
Gesamtanzahl des kinetischen Ausdrucks mit Zahl	33.819	27.154	1.948	1.794	72.656	3.978	35.873	4.356
Zahl und Enzym	24.234	19.818	1.531	1.298	40.146	2.475	20.427	2.682
Zahl und EC Nummer	8.900	5.619	603	408	8.627	559	2.945	740
Zahl, Enzym und EC Nummer	8.890	5.610	597	408	8.614	559	2.943	739
Zahl und Ligand	26.490	19.742	1.188	1.441	51.052	2.569	23.238	2.780
Zahl und Organismus	25.881	17.782	1.252	1.180	48.915	2.834	24.370	3.309
Zahl und Lokalisation	18.827	13.665	584	1.090	44.611	2.156	21.165	2.625
Zahl, Ligand und Organismus	20.423	13.032	776	940	34.600	1.833	15.735	2.131
Zahl, Enzym und Ligand	19.024	14.385	933	1.031	28.349	1.595	12.884	1.701
Zahl, EC Nummer und Ligand	7.227	4.265	366	336	6.046	375	2.048	473

	K_M	K_i	k_{cat}	$S_{0.5}$	IC_{50}	n_H	K_d	V_{max}
Zahl, Enzym, EC Nummer und Ligand	7.218	4.256	366	336	6.036	375	2.046	472
Zahl, Enzym und Organismus	18.714	13.109	999	879	26.968	1.836	14.308	2.041
Zahl, EC Nummer und Organismus	6.879	3.819	394	269	5.768	431	1.906	572
Zahl, EC Nummer, Enzym und Organismus	6.871	3.813	388	269	5.761	431	1.904	572
Zahl, Enzym und Lokalisation	13.413	9.903	464	814	24.376	1.352	12.221	1.575
Zahl, EC Nummer und Lokalisation	4.921	2.732	195	263	5.231	300	1.359	462
Zahl, Enzym, EC Nummer und Lokalisation	4.914	2.730	195	263	5.226	300	1.359	462
Zahl, Enzym, Ligand und Organismus	14.803	9.618	617	697	19.184	1.182	8.967	1.281
Zahl, EC Nummer, Ligand und Organismus	5.654	2.913	231	229	4.094	286	1327	357
Zahl, EC Nummer, Enzym, Ligand und Organismus	5.647	2.907	231	229	4.088	286	1.325	357
Zahl, Enzym, Ligand, Organismus und Lokalisation	8.429	5.480	200	469	12.593	678	5524	784
Zahl, EC Nummer, Ligand, Organismus und Lokalisation	3.211	1.621	72	158	2.751	145	654	233
Zahl, Enzym, EC Nummer, Ligand, Organismus und Lokalisation	3.206	1.621	72	158	2.750	145	654	233
Enzym und Ligand	37.830	20.536	4.691	1.668	38.299	2.138	21.872	13.263
EC Nummer und Ligand	14.596	6.149	1.818	567	8.246	480	3.666	4.122
Enzym, EC Nummer und Ligand	14.575	6.138	1.816	567	8.231	480	3.664	4.117
Enzym und Organismus	37.772	18.903	5.067	1.509	36.837	2.428	23.957	15.362
EC Nummer und Organismus	14.512	5.616	2.061	497	8.009	569	3.467	4.690
Enzym, EC Nummer und Organismus	14.497	5.608	2.052	497	7.996	569	3.465	4.686
Enzym und Lokalisation	27.183	14.321	2.192	1.302	33.296	1.827	20.419	12.304
EC Nummer und Lokalisation	10.166	3.983	895	421	7.274	401	2.454	3.561
Enzym, EC Nummer und Lokalisation	10.153	3.980	894	421	7.263	401	2.454	3.557

	k_{cat}/K_M	Spez. Aktivität	$t_{1/2}$	pI	K_a
Gesamtanzahl des kinetischen Ausdrucks	1.469	37.001	91.430	17.481	18.903
Gesamtanzahl des kinetischen Ausdrucks mit Zahl	71	3.396	45.129	10.158	3.246
Zahl und Enzym	41	2.517	11.502	5.200	1.180
Zahl und EC Nummer	12	1.131	3.117	1.900	221
Zahl, Enzym und EC Nummer	12	1.130	3.116	1.897	220
Zahl und Ligand	45	1.712	31.368	5.112	2.136
Zahl und Organismus	49	2.572	26.949	7.267	1.819
Zahl und Lokalisation	43	1.875	26.665	5.579	1.523
Zahl, Ligand und Organismus	27	1.267	18.735	3.618	1.198
Zahl, Enzym und Ligand	28	1.153	7.221	2.332	763
Zahl, EC Nummer und Ligand	8	499	1.948	826	151
Zahl, Enzym, EC Nummer und Ligand	8	498	1.947	823	151
Zahl, Enzym und Organismus	28	1.990	7.157	3.966	707
Zahl, EC Nummer und Organismus	10	916	1.986	1.514	138
Zahl, EC Nummer, Enzym und Organismus	10	916	1.986	1.511	137
Zahl, Enzym und Lokalisation	22	1.454	6.846	2.894	606
Zahl, EC Nummer und Lokalisation	4	681	1.820	1.064	83
Zahl, Enzym, EC Nummer und Lokalisation	4	680	1.819	1.064	83
Zahl, Enzym, Ligand und Organismus	15	894	4.496	1.752	462
Zahl, EC Nummer, Ligand und Organismus	6	396	1.210	647	88
Zahl, EC Nummer, Enzym, Ligand und Organismus	6	396	1.210	644	88
Zahl, Enzym, Ligand, Organismus und Lokalisation	5	545	2.821	1032	251

	k_{cat}/K_M	Spez. Aktivität	$t_{1/2}$	pI	K_a
Zahl, EC Nummer, Ligand, Organismus und Lokalisation	0	251	762	375	21
Zahl, Enzym, EC Nummer, Ligand, Organismus und Lokalisation	0	251	762	375	21
Enzym und Ligand	612	12.925	14.296	3.785	4.653
EC Nummer und Ligand	238	5.397	3.868	1.319	880
Enzym, EC Nummer und Ligand	237	5.384	3.866	1.316	877
Enzym und Organismus	569	18.064	1.3238	6.138	4.624
EC Nummer und Organismus	231	7.853	3.670	2.372	835
Enzym, EC Nummer und Organismus	231	7.841	3.670	2.367	829
Enzym und Lokalisation	382	14.573	13.030	4.523	3.947
EC Nummer und Lokalisation	145	6.324	3.489	1.666	643
Enzym, EC Nummer und Lokalisation	144	6.313	3.487	1.665	640

7.1.3. Arten der Verbindung

Für IC_{50}:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	21.606	24.901	2.387	19.003	4.596	0
Enzym	6.514	13.609	15.203	19.561	288	146
Organismus	4.715	12.988	27.837	23.821	570	0
Lokalisation	8.818	20.417	9.211	24.624	456	0
EC Nummer	39	45	884	11	0	11.135
Zahlenwert	0	0	0	0	0	0
pH	42.964	6.669	13	0	23.068	0
Temperatur	137	257	0	0	8	0

Für K_M:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	20.319	13.998	1.409	10.884	7.586	0
Enzym	5.627	10.891	11.100	22.646	284	729
Organismus	1.654	6.633	23.185	21.705	149	0
Lokalisation	2.841	9.145	10.540	16.481	430	0
EC Nummer	56	228	4.942	55	0	14.180
Zahlenwert	0	0	0	0	0	0
pH	1.9397	5.511	20	0	8.951	0
Temperatur	604	2.014	0	0	24	0

Für K_i:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	9.421	9.222	1.129	7.539	2215	0
Enzym	4.543	7.030	5.900	10.961	336	280
Organismus	1.821	4.122	11.749	8.909	146	0
Lokalisation	2.000	4.778	6.090	7.971	185	0
EC Nummer	28	122	1.823	40	2	6.352
Zahlenwert	17.676	1.918	22	0	7.594	0
pH	240	391	0	0	2	0

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Temperatur	76	343	0	0	8	0

Für kcat:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	1.079	2.615	182	2.127	286	0
Enzym	537	1.694	1.767	3.685	22	101
Organismus	112	628	3.098	2.526	10	0
Lokalisation	81	454	1.520	787	6	0
EC Nummer	4	33	561	14	0	2.421
Zahlenwert	1.302	367	1	0	282	0
pH	102	520	0	0	0	0
Temperatur	94	470	0	0	4	0

Für $S_{0.5}$:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	1.189	471	27	365	212	0
Enzym	154	501	436	1.029	2	21
Organismus	32	180	1.051	702	4	0
Lokalisation	114	370	406	809	15	0
EC Nummer	0	3	138	9	0	556
Zahlenwert	1.241	215	0	0	346	0
pH	30	84	0	0	0	0
Temperatur	5	34	0	0	0	0

Für n_H:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	594	1.467	98	1.116	183	0
Enzym	75	599	1.072	1.561	1	25
Organismus	27	302	2.121	1.326	3	0
Lokalisation	83	647	662	1.524	13	0
EC Nummer	0	3	197	4	0	542
Zahlenwert	2.888	539	0	0	551	0
pH	25	100	0	0	2	0
Temperatur	17	69	0	0	0	0

Für K_d:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	9.747	14.348	2.386	13.266	1.483	0
Enzym	2.340	6.982	11.380	14.223	61	101
Organismus	1.761	6.620	17.944	15.488	145	0
Lokalisation	3.891	10.817	6.426	15.386	478	0
EC Nummer	1	30	638	5	0	4.786
Zahlenwert	24.491	5.079	55	0	6.348	0

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
pH	267	792	0	0	6	0
Temperatur	528	1.391	0	0	28	0

Für V_{max}:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	37.60	9.389	1.152	7.399	846	0
Enzym	1.971	5.375	5.173	8.384	164	174
Organismus	528	3.565	11.956	9.677	70	0
Lokalisation	1.496	6.403	3.710	10.104	186	0
EC Nummer	8	56	1.109	13	0	5.314
Zahlenwert	2.381	1.125	7	0	851	0
pH	97	575	0	0	6	0
Temperatur	119	713	0	0	6	0

Für k_{cat}/K_M:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	191	351	32	296	101	0
Enzym	76	208	225	372	4	6
Organismus	34	153	345	423	16	0
Lokalisation	57	172	199	258	11	0
EC Nummer	0	2	60	1	0	289
Zahlenwert	30	13	1	0	31	0
pH	0	0	0	0	0	0
Temperatur	9	25	0	0	0	0

Für die spezifische Aktivität:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	4.409	7.967	1.898	6.555	910	0
Enzym	6.339	4.725	3.890	6.565	1.810	229
Organismus	1.161	5.338	9.885	10.518	109	0
Lokalisation	3.689	6.870	3.737	7.329	748	0
EC Nummer	135	210	1.693	19	0	7.915
Zahlenwert	2.396	607	3	0	396	0
pH	90	405	0	0	2	0
Temperatur	128	615	0	0	2	0

Für $t_{1/2}$:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	14.169	19.919	4.803	22.777	1.379	0
Enzym	1.869	5.142	9.518	5.575	73	75
Organismus	2.108	9.324	23.382	16.952	311	0
Lokalisation	11.091	19.587	6.572	15.093	681	0
EC Nummer	3	21	510	8	0	5.462

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Zahlenwert	26.606	9.210	42	0	9.513	0
pH	314	779	0	0	8	0
Temperatur	1.206	1.698	0	0	57	0

Für pI:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	951	3.699	1.001	3.217	101	0
Enzym	669	2.076	2.281	3.008	41	138
Organismus	254	1.694	4.457	5.412	37	0
Lokalisation	405	2.163	2.748	4.053	56	0
EC Nummer	5	37	927	17	0	2.011
Zahlenwert	4.860	2.276	0	0	3.022	0
pH	1.153	645	0	0	57	0
Temperatur	19	227	0	0	0	0

Für K_a:

	direkt	indirekt aus Satz	indirekt aus Abstract	indirekt aus Titel	Auflistung	automatisch
Ligand	2.292	4.442	782	3.998	464	0
Enzym	433	2.010	2475	2.193	39	34
Organismus	371	1.942	4.758	3.690	25	0
Lokalisation	583	3.022	1.958	3.927	69	0
EC Nummer	4	22	166	0	0	1.097
Zahlenwert	1.966	814	0	0	466	0
pH	55	285	0	0	2	0
Temperatur	161	473	0	0	2	0

7.1.4. Verteilung von extrahierten kinetischen Zahlenwerten

Für Kategorien mit Konzentrationsangabe [mM]:

Intervall	K_M Anzahl [%]	K_i Anzahl [%]	$S_{0.5}$ Anzahl [%]	IC_{50} Anzahl [%]	K_d Anzahl [%]
]e-8; e-7]	2	30	0	63	39
]e-7; e-6]	7	301	3	457	776
]e-6; e-5]	78	1.469	16	2.490	4.731
]e-5; e-4]	257	2.696	69	5.737	6.336
]e-4; e-3]	791	3.011	118	8.296	4.474
]e-3; e-2]	2.325	3.367	213	9.717	3.543
]e-2; e-1]	5.128	3.835	221	12.018	2.628
]e-1; e0]	8.255	4.061	305	11.512	2.013
]e0; e+1]	7.373	2.707	286	4.532	1.304
]e+1; e+2]	4.424	1.516	237	1.542	752
]e+2; e+3]	1.673	532	144	611	363

Für die anderen Kategorien:

k_{cat} [s^{-1}]		K_a [mM^{-1}]		Spez. Aktivität [U/mg]		$t_{1/2}$ [s]	
Intervall	Anzahl [%]	Intervall	Anzahl [%]	Intervall	Anzahl [%]	Intervall	Anzahl [%]
]e-9; e-8]	1	]e-2; e-1]	1	]e-9; e-8]	1	]e-9; e-8]	30
]e-8; e-7]	0	]e-1; e0]	0	]e-8; e-7]	2	]e-8; e-7]	25
]e-7; e-6]	3	]e0; e+1]	0	]e-7; e-6]	3	]e-7; e-6]	25
]e-6; e-5]	2	]e+1; e+2]	1	]e-6; e-5]	1	]e-6; e-5]	16
]e-5; e-4]	14	]e+2; e+3]	2	]e-5; e-4]	5	]e-5; e-4]	19
]e-4; e-3]	43	]e+3; e+4]	16	]e-4; e-3]	18	]e-4; e-3]	40
]e-3; e-2]	99	]e+4; e+5]	35	]e-3; e-2]	20	]e-3; e-2]	71
]e-2; e-1]	147	]e+5; e+6]	46	]e-2; e-1]	25	]e-2; e-1]	119
]e-1; e0]	256	]e+6; e+7]	62	]e-1; e0]	71	]e-1; e0]	120
]e0; e+1]	468	]e+7; e+8]	77	]e0; e+1]	220	]e0; e+1]	216
]e+1; e+2]	494	]e+8; e+9]	81	]e+1; e+2]	334	]e+1; e+2]	1.105
]e+2; e+3]	287	]e+9; e+10]	64	]e+2; e+3]	274	]e+2; e+3]	5.322
]e+3; e+4]	63	]e+10; e+11]	59	]e+3; e+4]	123	]e+3; e+4]	13.927
]e+4; e+5]	12	]e+11; e+12]	38	]e+4; e+5]	21	]e+4; e+5]	14.243
]e+5; e+6]	12	]e+12; e+13]	39	]e+5; e+6]	18	]e+5; e+6]	5.720
]e+6; e+7]	7	]e+13; e+14]	21	]e+6; e+7]	20	]e+6; e+7]	1.904
]e+7; e+8]	3	]e+14; e+15]	7	]e+7; e+8]	41	]e+7; e+8]	409
]e+8; e+9]	0	]e+15; e+16]	1	]e+8; e+9]	23	]e+8; e+9]	368
]e-9; e-8]	1	]e+16; e+17]	0			]e+8; e+10]	61
]e-8; e-7]	0					]e+10; e+11]	21

7.1.5. Verteilung von extrahierten Organismen

Tiergruppe	Anzahl
Tiere	212.929
Säugetiere	197.400
Bakterien	28.893
Pflanzen	14.988
Pilze	11.253
Viren	8.538

7.1.6. Verteilung der Publikationsdaten

Jahr	Ges. Anzahl Artikel	K_M	K_i	k_{cat}	$S_{0.5}$	IC_{50}	K_d	V_{max}	n_H	k_{cat}/K_M	Spez. Aktivität	$t_{1/2}$	pI	K_a
1975	198.023	1.522	429	67	33	88	487	463	76	3	1075	1.000	310	255
1976	199.338	1.560	512	86	39	176	566	463	100	5	1140	1.091	316	304
1977	204.684	1.136	367	46	25	163	413	356	65	3	901	1.082	304	189
1978	213.111	1.152	332	49	18	160	428	348	40	1	818	1.077	323	229
1979	221.029	1.274	366	57	26	203	550	440	45	4	813	1.253	388	287
1980	220.550	1.533	463	91	37	317	774	526	68	7	1.033	1.300	426	359
1981	221.560	1.636	566	78	83	419	934	654	93	10	1.067	1.765	447	406
1982	229.385	1.891	685	88	53	687	1.143	814	91	11	1.294	1.995	594	395
1983	239.835	2.244	873	145	104	987	1.650	1.023	184	20	1.339	2.361	787	540
1984	245.182	2.301	948	134	123	1.258	1.885	1.117	180	27	1.398	2.459	773	526
1985	254.487	2.503	1.158	156	112	1.595	2.096	1.240	188	29	1.371	2.729	706	604
1986	263.984	2.554	1.258	190	124	1.944	2.300	1.378	173	21	1.447	3.008	751	586
1987	287.112	2.435	1.309	235	139	2.034	2.273	1.496	178	24	1.317	3.087	627	584

Jahr	Ges. Anzahl Artikel	K_M	K_i	k_{cat}	$S_{0.5}$	IC_{50}	K_d	V_{max}	n_H	k_{cat}/K_M	Spez. Aktivität	$t_{1/2}$	pI	K_a
1988	290.332	2.553	1.270	223	127	2.439	2.590	1.496	223	43	1.311	3.203	702	558
1989	300.253	2.491	1.480	272	144	2.922	2.788	1.516	193	46	1.385	3.247	654	643
1990	305.650	2.787	1.551	313	139	3.137	2.970	1.678	228	55	1.282	3.496	700	663
1991	304.513	2.634	1.542	334	111	3.651	2.796	1.555	216	48	1.305	3.331	657	614
1992	306.961	2.702	1.747	402	189	3.992	2.983	1.582	220	43	1.246	3.304	612	602
1993	307.144	2.773	1.725	396	120	4.522	2.892	1.499	249	55	1.146	3.277	597	612
1994	311.876	2.657	1.860	401	125	5.058	3.014	1.431	246	68	1.079	3.224	552	598
1995	319.306	2.766	1.852	487	119	5.166	2.793	1.371	270	49	1.078	3.270	582	619
1996	325.867	2.786	1.803	515	98	5.347	2.767	1.472	249	82	1.074	3.392	441	600
1997	330.490	2.563	1.839	506	77	5.171	2.672	1.389	202	102	991	3.137	448	672
1998	341.615	2.440	1.777	537	74	5.315	2.335	1.267	218	64	970	3.093	467	735
1999	352.448	2.478	1.840	530	85	5.242	2.344	1.203	212	73	912	3.144	439	758
2000	377.506	2.430	1.725	508	92	5.069	2.158	1.185	194	85	922	3.122	410	728
2001	379.628	2.331	1.535	552	56	5.035	2.016	1.013	182	77	816	2.995	420	655
2002	391.107	2.148	1.496	570	87	5.061	1.911	974	145	61	821	2.916	404	693
2003	404.360	2.067	1.466	536	61	4.938	1.956	1.053	143	58	779	2.812	361	599
2004	421.283	2.056	1.526	556	67	5.936	1.970	959	148	73	824	2.984	441	690
2005	433.584	1.919	1.530	508	72	5.616	2.142	879	139	56	777	3.059	451	760
2006	471.804	1.845	1.416	479	54	6.121	1.864	830	118	52	789	2.990	496	765

7.1.7. Precision und Recall

In 1002 validierten Einträgen enthaltene Informationen zu verschiedene Kategorien:

	in Abstract enthalten	richtig	falsch
Temperatur	39	16	4
pH	40	16	13
Lokalisation	767	492	44
Ligand	938	496	205
Kinetischer Zahlenwert	697	538	48
EC Nummer	58	31	3
Enzym	686	457	58
Organismus	886	657	27

Für den gesamten Eintrag:

	richtig	Gesamtanzahl der Einträge
K_M	872	1.192
k_{cat}/K_M	16	48
Spez. Aktivität	368	552
$S_{0.5}$	32	32
K_a	24	136
IC_{50}	872	1.584
n_H	88	112
K_d	640	912
$t_{1/2}$	1.032	1.616
k_{cat}	128	168
pI	120	240
K_i	552	816
V_{max}	440	584

7.2. Tabellenverzeichnis

7.3. Abbildungsverzeichnis

Entitäten verdeutlicht; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren zusätzlich verbindende Ausdrücke; rote Kreuze beenden die Verbindung.

Abbildung 14. Schema der inhaltlichen direkten Verbindung anhand eines Beispielsatzes mit zwei Auflistung für den Sonderfall der Einheiten; identifizierte Entitäten sind mit einem grünen Kreis unterlegt; blau, rosa und violett unterlegte Entitäten sind Teil der Auflistung; die jeweilige Verbindung von den Werten in den Listen untereinander ist durch die gleiche farbliche Unterlegung der Entitäten verdeutlicht; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren zusätzlich verbindende Ausdrücke; rote Kreuze beenden die Verbindung.

Abbildung 15. Schema der inhaltlichen indirekten Verbindung anhand eines Beispielsatzes; identifizierte Entitäten sind grün oder blau unterlegt; grüne und blaue Pfeile verbinden Entitäten inhaltlich; blaue Pfeile markieren verbindende Ausdrücke; rote Kreuze beenden die Verbindung bei Kommata, Satzende oder fehlenden verbindenden Ausdrücken; die blau unterlegte Entität wurde als einzige Entität dieser Kategorie der inhaltlichen Verbindung hinzugefügt ohne direkte Verbindung hinzugefügt.

Abbildung 16. Tabelle der Suchergebnisse mit Verknüpfungen (unterstrichen und rot unterlegt) zu einer ausführlicheren Präsentation der Ergebnisse.

Abbildung 17. Ausführliche Darstellung der Ergebnisse.

Abbildung 18. Gesamtanzahl der extrahierten Werte unterschieden nach verbundenem kinetischem Wert und Kategorie Teil 1 (siehe Anhang Punkt 7.1.1).

Abbildung 19. Gesamtanzahl der extrahierten Werte unterschieden nach verbundenem kinetischem Wert und Kategorie Teil 2 (siehe Anhang Punkt 7.1.1).

Abbildung 20. Prozentuale Verteilung der Kombinationen verschiedener Kategorien bezogen auf die Anzahl der jeweiligen kinetischen Kategorie Teil 1. Zahl stellt eine Abkürzung für einen Zahlenwert dar; Enzym ist eine Abkürzung für Enzyminformation und umfasst sowohl Enzymnamen als auch EC Nummern (siehe Anhang Punkt 7.1.2).

Abbildung 21. Prozentuale Verteilung der Kombinationen verschiedener Kategorien bezogen auf die Anzahl der jeweiligen kinetischen Kategorie Teil 2. Zahl stellt eine Abkürzung für einen Zahlenwert dar; Enzym ist eine Abkürzung für Enzyminformation und umfasst sowohl Enzymnamen als auch EC Nummern (siehe Anhang Punkt 7.1.2).

Abbildung 34. Menge an Abstracts bezogen auf den beinhalteten kinetischen Ausdruck und die Menge an veröffentlichten Abstract im Erscheinungsjahr Teil 1 (siehe Anhang Punkt 7.1.6).

Abbildung 35. Menge an Abstracts bezogen auf den beinhalteten kinetischen Ausdruck und die Menge an veröffentlichten Abstract im Erscheinungsjahr Teil 2 (siehe Anhang Punkt 7.1.6).

Abbildung 36. Graphische Darstellung des Recalls unterteilt nach Kategorien (siehe Anhang Punkt 7.1.7).

Abbildung 37. Graphische Darstellung der Precision-Werte unterteilt nach Kategorien (siehe Anhang Punkt 7.1.7).

Abbildung 38. Genauigkeit der gesamten Einträge unterschieden nach den beinhalteten kinetischen Ausdrücken. 100% entspricht 149 (K_M), 6 (k_{cat}/K_M), 69 (spezifische Aktivität), 4 ($S_{0.5}$), 17 (K_a), 198 (IC_{50}), 14 (n_H), 114 (K_d), 202 ($t_{1/2}$), 21 (k_{cat}), 30 (pI), 102 (K_i) und 73 (V_{max}) (siehe Anhang Punkt 7.1.7).

8. 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, Karten 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 unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. D. Schomburg betreut worden.

Stephanie Heinen

9. Lebenslauf

Name: Stephanie Eva Heinen

Geburtstag: 11.03.1980

Geburtsort: Köln

Familienstand: ledig

Studium:

seit Dez. 2005:	Doktorarbeit bei Prof. Dr. D. Schomburg im Bereich theoretische Biochemie (Bioinformatik) mit dem Titel „Entwicklung einer Textminingmethode zur automatisierten Extraktion von kinetischen Informationen aus der Literatur“
bis 29. Nov. 2005:	Diplomarbeit in der theoretischen Biochemie (Bioinformatik) zum Thema „Textmining-Programm zur Automatisierung der Suche nach kinetischen Werten in der PubMed“ bei Prof. Dr. D. Schomburg Note des Diploms: „sehr gut“
Nov. bis Dez. 2004:	Mündliche Diplomprüfung mit den folgenden Ergebnissen: 1. Hauptfach (Biochemie): „sehr gut“ 1. Nebenfach (Entwicklungsbiologie): „sehr gut -“ 2. Nebenfach (Tierphysiologie): „sehr gut -“
26. Sep. 2002:	Vordiplom mit der Note „sehr gut“
Okt. 2000:	Beginn des Studiums für Biologie und Geschichte auf Sek. 2. Im gleichen Semester Wechsel zum Studiengang Diplom Biologie.

Schulbildung:

1990 bis 1999: Besuch des Montessori Gymnasiums in Köln Bickendorf mit Abitur (Abiturnote 1,4)

1986 bis 1990: Besuch der Montessori Grundschule in Köln Bickendorf.