Herausforderungen bei der Extraktion von biochemischen Daten aus der Literatur

Ulrike Wittig, Renate Kania, Isabel Rojas, Wolfgang Müller

Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg
Email: Ulrike.Wittig@h-its.org

Abstract: Biochemische Daten in der wissenschaftlichen Literatur liegen in einem nur wenig strukturierten und standardisierten Format vor. Um diese Informationen nutzen und automatisieren zu können, entstand eine Vielzahl von Datenbanken, für die publizierte Daten größtenteils manuell aus der Literatur extrahiert werden, um sie Biowissenschaftlern zur Nutzung zur Verfügung zu stellen. Die Herausforderung bei dem Betrieb solcher Datenbanken besteht unter anderem in der Sicherung der Qualität der Daten. Dies bedeutet, dass viel Zeit von biologischen Experten investiert werden muss, um die Daten aus der Literatur zu extrahieren und für die Eingabe in die Datenbank vorzubereiten, um sie bestehenden Standards anzupassen. Dies erzeugt einen Großteil der zum Betrieb erforderlichen Kosten und beeinflusst damit direkt die Machbarkeit von Projekten. In dieser Publikation beschreiben wir anhand der Datenbank SABIO-RK, welche Probleme von eventuellen automatischen Methoden gelöst werden müssten, um menschliche Arbeitskraft zu ersetzen.

1 Einleitung

Obwohl eine Vielzahl wissenschaftlicher Publikationen in den Lebenswissenschaften heutzutage elektronisch erreichbar ist, ist die Art, in der Informationen publiziert werden, immer noch traditionell innerhalb eines Textes, in Tabellen oder als Abbildungen. Es gibt nur wenige Zeitschriften, die von ihren Autoren die Publikation von Ergebnissen in strukturierter und standardisierter Form verlangen. Um Wissenschaftler bei der Suche nach relevanten Daten für ihre Arbeit zu unterstützen, arbeiten mehrere Gruppen an der Entwicklung von wissenschaftlichen Datenbanken, die Literaturdaten in suchbarer und strukturierter Form enthalten. Allerdings tun sie dies zu recht hohen Kosten, da menschliche Arbeitskraft nötig ist, um die benötigte Information in hoher Qualität herauszufiltern. In dieser Publikation geben wir einen Überblick über einige Probleme, die es zu lösen gilt, um die zur Zeit von Menschen durchgeführten Prozesse wirksam mit Rechnern zu unterstützen. Wir werden am Beispiel der SABIO-RK Datenbank, die in unserer Gruppe entwickelt wird, die Herausforderungen bei der Extraktion von Daten aus der Literatur aufzeigen. SABIO-RK (http://sabio.villa-bosch.de/SABIORK) [1,2], ist eine kuratierte, online zugängliche Datenbank für biochemische Reaktionen und deren detaillierten Informationen zu Reaktionsgeschwindigkeiten. Derzeit wird die Mehrzahl der notwendigen Daten manuell aus der Literatur extrahiert. Neue experimentelle Daten können auch direkt aus dem Experiment heraus eingepflegt werden. Die für die Eingabe in die SABIO-RK Datenbank wichtigen Daten werden dabei aus der wissenschaftlichen

Literatur der letzten 50 Jahre gesammelt. SABIO-RK-Datenbanknutzer sind überwiegend Biowissenschaftler, die diese so genannten enzymkinetischen Daten zur Erstellung von Computermodellen und zur Simulation von biochemischen Reaktionen, Stoffwechselwegen und komplexen biologischen Netzwerken verwenden. Die Enzymkinetik beschreibt dabei die Konzentrationsabhängigkeit der Reaktionsgeschwindigkeit einer biochemischen Reaktion und bestimmt Parameter für ein einzelnes Protein, welches als Enzym, einem so genannten Biokatalysator fungiert. Da Enzyme biochemische Reaktionen beschleunigen und kontrollieren, ist die enzymkinetische Analyse zum Verständnis von Enzymfunktionen unerlässlich. Die relevanten Daten liegen aber in der Literatur in einer Form vor, die die Weiterverwendung (z.B. zur Modellbildung) erschwert. SABIO-RK ist nun ein Dienst, der das Auffinden der wichtigen Parameter und somit die Nachnutzung der Daten erleichtert.

2 Problemstellung

In diesem Abschnitt geben wir einen Überblick über die wichtigsten Probleme, mit denen wir bei der Extraktion von relevanten Daten für SABIO-RK konfrontiert werden. Diese Probleme treten aber nicht nur exklusiv bei SABIO-RK auf, sondern lassen sich auch auf viele andere wissenschaftliche Datenbanken übertragen. Um die Probleme darzustellen, haben wir in Abbildung 1A vier Seiten einer insgesamt 7-seitigen Publikation ausgewählt. Ohne detaillierte Textstellen zu analysieren, soll die Abbildung anhand der Farbmarkierung die räumliche Verteilung verschiedener Daten der gleichen Entität innerhalb einer Publikation aufzeigen. Abbildung 1B zeigt demgegenüber die strukturierte Darstellung eines SABIO-RK Eintrages in der webbasierten Benutzeroberfläche der Datenbank, der die extrahierten Informationen aus der Publikation aus Abbildung 1A enthält, Aus dieser Publikation resultieren insgesamt 6 unterschiedliche Einträge in der SABIO-RK Datenbank.

Räumliche Verteilung in der Publikation: Publikationen experimenteller biologischer Daten folgen in der Regel einem klassischen Schema, bei dem nach einer Einleitung, die das Experiment beschreibt, also Metadaten und Hintergrundinformationen enthält, der Methodenteil mit Angaben über Versuchsabläufe und Messmethoden anschließt. Darauf folgt die Auflistung der eigentlichen Ergebnisse, die dann wiederum in einem vom Ergebnisteil unabhängigen getrennten Abschnitt interpretiert und diskutiert werden. Diese strikte Trennung von Informationen innerhalb einer Publikation bedingt Probleme bei einer möglichen automatischen Zuordnung von experimentellen Ergebnissen. Zusätzlich werden experimentelle Ergebnisse entweder im Fließtext beschrieben und können weit über den Text verstreut vorkommen oder sie werden in Tabellen oder Graphiken dargestellt.

Wenn im Methodenteil z.B. beschrieben wird, dass die Parameter durch die Anwendung einer Gleichung bestimmt wurden, die den kinetischen Mechanismus der ausgewählten Reaktion beschreibt, ist es wichtig, alle zur Gleichung gehörenden Parameter zu finden und einzufügen. In unserem Beispiel enthalten die *grün* markierten Textstellen kinetische Parameter und Informationen über mathematische Gleichungen, mit denen die

Parameter kalkuliert wurden. Diese befinden sich in dieser Beispielpublikation im Text über die gesamte Publikation verteilt, aber auch in einer Tabelle und in einer graphischen Darstellung. Dass, wie in unserem Beispiel, nicht alle Parameter einer Gleichung zusammen in einer Tabelle dargestellt werden, ist leider kein Einzelfall.

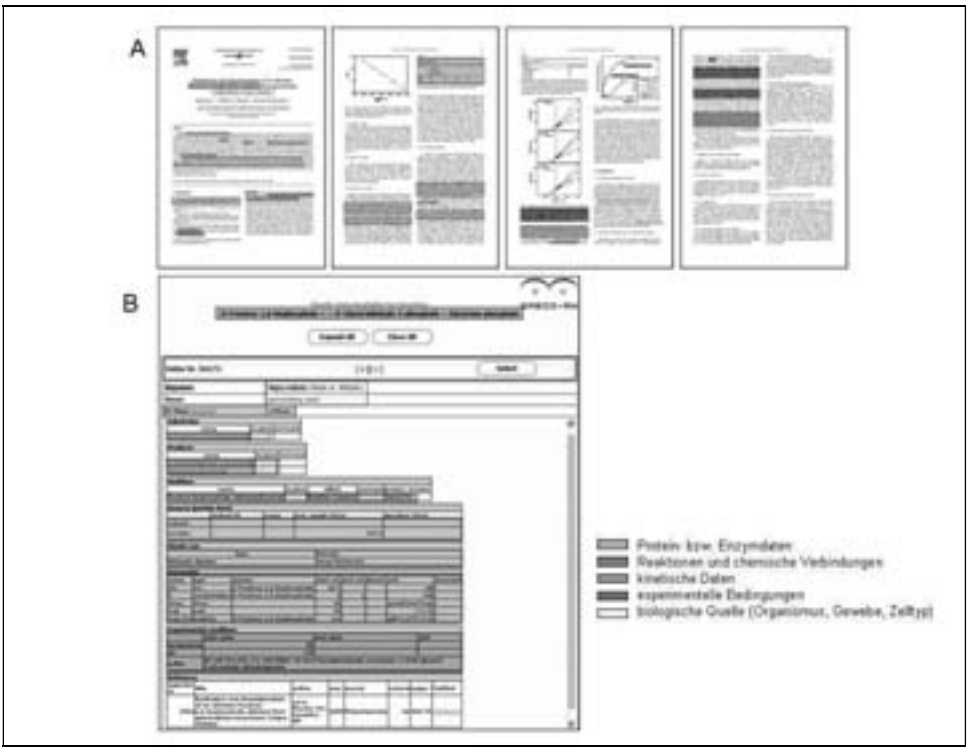

Abbildung 1: Farbmarkierung relevanter Daten in einer Publikation für die Eingabe in SABIO-RK (A) und deren strukturierte Darstellung in einem Eintrag im SABIO-RK Webinterface (B)

Fehlende Daten: Ein Teil der Unvollständigkeit der Daten beruht auf dem Alter der Artikel, z.B. fehlen nähere Angaben zu Proteinen/Enzymen, da zum Zeitpunkt der Publikation nicht bekannt war, dass ein bestimmtes Enzym als Isoform vorliegen kann. Dies bedeutet, ein Organismus besitzt ein oder mehrere Proteine, die der gleichen Enzymklasse angehören und somit die gleiche biochemische Reaktion katalysieren. Diese Proteine unterscheiden sich aber z.B. in der Aminosäurezusammensetzung und besitzen somit andere Molekulargewichte. In ca. 85% der Publikationen gibt es keine Referenzen zu Einträgen von Proteinsequenzen in der Standard-Proteindatenbank UniProt [4]. Diese Datenbank wurde Ende der 1980iger Jahre entwickelt und gilt heute als Standardreferenz für Proteindaten. Aber auch die Mehrzahl der neueren Publikationen enthalten keine Referenzen zu UniProt. Hier sollten die Autoren von

Seiten der wissenschaftlichen Zeitschriften mehr angehalten werden, solch wichtige Referenzen anzugeben.

Biochemische Reaktionen werden durch chemische Verbindungen beschrieben, die als Ausgangssubstrate, Endprodukte und Modifikatoren fungieren. Modifikatoren sind Verbindungen, die die Reaktionsgeschwindigkeit einer Reaktion beeinflussen können, wie z.B. Katalysatoren (Enzyme), Inhibitoren oder Aktivatoren. Für die Eingabe von Reaktionen in SABIO-RK fehlen allerdings in ca. 14% der Publikationen die vollständigen Reaktionsgleichungen. Es werden häufig nur die Substrate genannt, die für die Messungen verwendet wurden und eine Angabe von Reaktionsprodukten fehlt.

Die Wirkungsweise eines Enzyms ist stark abhängig von Temperatur und pH-Wert. Ein großer Vorteil von SABIO-RK gegenüber anderen Enzymdatenquellen ist daher nicht nur die Darstellung der biochemischen Reaktionen und ihrer Reaktionsgeschwindigkeits-relevanten Daten, sondern auch die zusätzlich gespeicherte Information zu den experimentellen Bedingungen wie Temperatur, pH-Wert und Pufferzusammensetzung, unter denen die Daten gemessen wurden. Aber auch hier sind wir mit dem Fehlen oder der Ungenauigkeit von wichtigen Angaben konfrontiert In. 12% der Publikationen gibt es überhaupt keine Angabe zur Temperatur und, bei etwa 6% steht die nicht sehr präzise Angabe Zimmertemperatur (room temperature). Eine mit etwa 10% des Öfteren praktizierte Art der Angabe von experimentellen Bedingungen, ist der Verweis auf eine andere Publikation in der Liste des Literaturverzeichnisses. In diesen Fällen ist eine zeitaufwändige Recherche in den Referenzpublikationen und zum Teil in weiterführenden Referenzen erforderlich. In ca. 20% der Publikationen werden die zu den experimentellen Bedingungen gehörenden Zusammensetzungen des Puffers und die Stoffmengenangaben für Reaktionszusätze nicht in mol/l als Standardeinheit angeben sondern müssen erst manuell auf ein Standardvolumen umgerechnet werden.

Inkonsistenzen: Experimentelle Bedingungen im Methodenteil eines Artikels müssen nicht zwangsläufig übereinstimmen mit den experimentellen Bedingungen in der Legende von Tabellen oder Graphiken, in denen kinetische Parameter angegeben werden. Dies trifft auf etwa 6% der Publikationen zu. Darüber hinaus werden häufig im Methodenteil mehrere verschiedene experimentelle Bedingungen angegeben. Ein Beispiel hierfür ist, dass der Km-Wert unter anderen Bedingungen und mit einer anderen Methode bestimmt wurde als der Kd-Wert. Beide Werte können aber in einer Tabelle angegeben sein. Dann erfordert eine automatische Zuordnung der Methoden die Trennung der Parameter in der Tabelle. Auch innerhalb von SABIO-RK kommt es zur Trennung dieser Parameter und es werden aus diesen Daten unterschiedliche Einträge generiert, da unterschiedliche experimentelle Bedingungen vorliegen.

Vergleiche über Artikel hinweg: Teilweise werden Parameter aus einem ausgewählten Artikel verglichen mit Parametern, die aus der Literatur stammen. Diese werden dann in die Ergebnistabellen mit aufgenommen und nur in der Legende der Tabelle als Referenzwerte ausgewiesen. Hier besteht die Herausforderung darin, die Referenzwerte herauszufiltern und der entsprechenden referenzierten Publikation zuzuordnen.

Die ausgewählten Beispiele zeigen die Komplexität der Problematik und verdeutlichen die Anforderungen, die notwendig sind, um eine sinnvolle manuelle und/oder automatische Informationsextraktion biochemischer Daten aus der Literatur durchführen zu können.

3 Lösungsmöglichkeiten

Die oben genannten Beispiele für Probleme bei der Extraktion von Daten aus der Literatur zeigen, dass derzeit sehr viel manuelle Arbeit von Seiten der Biowissenschaftler bei der Eingabe der Daten in SABIO-RK notwendig ist. Weit verbreitet sind Software-Anwendungen, die die fehlerarme Zuordnung von Namen zu Einträgen in Ontologien erleichtern. Die Möglichkeiten reichen von Auswahllisten zu der linguistischen Analyse von Entitätennamen [5] zu Ansätzen, Nutzer mittels Machine Learning zu unterstützen [6,7].

Für *zukünftige* Publikationen sind in erster Linie die Autoren gefordert, ihre Daten vollständig und standardisiert in der Literatur zu beschreiben. Hierfür sollten die Zeitschriftenverlage idealerweise Formate und Standards vorgeben oder standardisierte Eingabemöglichkeiten anbieten. In Zukunft sind die Autoren der Publikationen gefordert, sich an vorgegebene Standards zu halten. Aber auch die Gutachter und Herausgeber der Zeitschriften sollten darauf achten, dass alle notwendigen Informationen angegeben wurden. Ideal wäre eine enge Zusammenarbeit von Datenbanken und Herausgebern um z.B. eine Eingabemaske bereitzustellen und die Autoren anzuhalten, ihre Daten an die entsprechende Datenbank weiterzuleiten. Ein Schritt in diese Richtung wird von der STRENDA Initiative [8] unternommen, die eine Liste von Anforderungen mit Vorschlägen für die Publikation von Enzymdaten erstellt hat. Diese Liste enthält ein Minimum an Informationen, das notwendig ist, um ein Enzym und seine relevanten Daten vollständig zu charakterisieren. Einige wissenschaftliche Zeitschriften haben bereits zugestimmt, ihren Autoren die Anforderungsliste der STRENDA Initiative als Grundlage für das Publizieren von Enzymdaten nahe zu legen. Auch die Gründung der International Society for Biocuration im Jahre 2009 [9] zeigt die Wichtigkeit der Sicherung der Datenqualität und die Notwendigkeit der Definition von Standards durch die stetig wachsende Anzahl von Daten in den Biowissenschaften aufgrund der Anwendung immer modernerer Meßmethoden [10].

Aber auch wenn mehr und mehr Daten neu sind, bleibt immer noch die Erschließung bereits erfolgter Messungen für die Speicherung in Datenbanken, also *Legacy-Publikationen.* Bei den im Abschnitt Problemstellung genannten Beispielen ergibt sich die Frage, inwieweit es möglich ist, die Extraktion der Informationen aus der Literatur durch Nutzung von Softwareanwendungen zu automatisieren. Für einen Großteil der fehlenden und ungenauen Angaben in der Literatur gibt es kaum eine Möglichkeit, die zusätzlichen Informationen automatisch zu generieren.

Dagegen wäre bei Problemen der Zuordnung von Daten und von über den gesamten Text der Publikation verteilten Informationen eine automatische Lösung zumindest ein

Ziel. Am Beispiel der Extraktion von kinetischen Daten aus der Literatur besteht eine Herausforderung darin, die richtigen experimentellen Bedingungen den kinetischen Parametern zuzuordnen, fehlende Informationen aus der Referenzliteratur zu extrahieren oder Zusammenhänge zwischen Parametern herzustellen. Letzteres bezieht sich z.B. auf die Analyse, welche Parameter unter identischen experimentellen Bedingungen bestimmt wurden und somit gemeinsam in einer mathematischen Gleichung repräsentiert werden können. Ein Beispiel hierfür ist die Suche nach experimentellen Ergebnissen, die z.B. als Parameterwerte Km und Vmax angegeben werden, die beide gemeinsam in der mathematischen Gleichung für die Michaelis-Menten-Kinetik dargestellt werden können. In über 90%der Publikationen werden die Ergebnisse im Text nicht zusammen mit der mathematischen Gleichung und den experimentellen Bedingungen, unter denen sie ermittelt wurden, angegeben. Stand der Technik ist die automatische Markierung von relevanten Informationen in einer Publikation. Aus unserer Sicht muss diese mit Methoden komplementiert werden, wenige mögliche Zuordnungen von Daten unterschiedlicher Entitäten vorzuschlagen. Diese Kombination von Markierungen und möglichen Zuordnungen würde die Wissenschaftler bei der Suche und Zuordnung von Daten unterstützen und die Extraktion von Daten erheblich erleichtern.

Danksagung

Das SABIO-RK Projekt wird durch die Klaus-Tschira-Stiftung und das Bundesministerium für Bildung und Forschung (SysMO/ Virtuelle Leber) finanziert. Die Autoren danken Henriette Slogsnat für anregende Diskussionen.

Literaturverzeichnis

[1] Wittig U *et al.* (2006) SABIO-RK: integration and curation of reaction kinetics data. Lecture Notes in Computer Science, 4075, 94-103

[2] Rojas I *et al.* (2007).Storing and annotating of kinetic data. In Silico Biol., 7 (Suppl 2) (17822389):37-44

[3] Hucka M *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 19, 524-31

[4] The UniProt Consortium. (2008) The Universal Protein Resource (UniProt)**.** Nucleic Acids Res. 36:D190-D195

[5] Engelken H *et al.* (2009) Flache und semantische Verarbeitung von Namen biochemischer Verbindungen. In INFORMATIK 2009 - Im Focus das Leben, Beiträge der 39. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Lübeck, Germany, LNI

[6] Van Auken K *et al*. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. BMC Bioinformatics. 10:228

[7] Doms A, Schroeder M (2005) GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 33:W783-6

[8] STRENDA: http://strenda.org/

[9] International Society for Biocuration: http://biocurator.org/

[10] Howe D *et al.* (2008) Big data: The future of biocuration, Nature. 455(7209):47-50.