René Ranzinger
Dr. sc. hum.

GlycomeDB: Integration von Kohlenhydratstrukturdatenbanken

Geboren am 20.09.1976 in Zittau
Diplom/Master der Fachrichtung Informatik am 03.11.2004 an der Fachhochschule Darmstadt

Promotionsfach: Medizinische Biometrie u. Informatik
Doktorvater: Prof. Dr. rer. nat. Thomas Wetter

Das größte Problem bei der Arbeit mit Datenbanken für Kohlenhydratstrukturen ist die Tatsache, dass die Strukturen und dazu gespeicherte Informationen, wie beispielsweise taxonomische Zuordnungen oder experimentelle Daten, über zahlreiche Datenbanken weltweit verteilt sind. Dabei haben die einzelnen Datenbanken einen überlappenden Datenbestand. Erschwerend kommt hinzu, dass für jede Datenbank ein eigenes Benennungsschema für Monosaccharide und ein eigenes Sequenzformat für Kohlenhydrate entwickelt wurde. Dies führt dazu, dass jede der Datenbanken eine isolierte Insel von Informationen ist, welche mit den anderen Inseln nicht in Verbindung steht und keine Informationen austauscht.

Das GlycomeDB Projekt wurde gestartet, um dieser Isolation und Verteilung der Informationen entgegenzuwirken. Im Rahmen des Projektes wurden die Strukturen und die taxonomischen Zuordnungen der sieben größten frei verfügbaren Datenbanken für Kohlenhydratstrukturen in einer neuen Datenbank mit dem Namen GlycomeDB zusammengeführt. Diese sieben weltweit verteilten Datenbanken sind die BCSDB, die CarbBank, die CFG Datenbank, die GLYCOSCIENCES.de, die GlycoBase (Dublin), die GlycoBase (Lille) und die KEGG Glycan Datenbank. Der wichtigste Arbeitsschritt bei der Erstellung der GlycomeDB war die Vereinheitlichung der zu speichernden Informationen. Als Speicherformat für die Kohlenhydratstrukturen wurde das Sequenzformat GlycoCT verwendet und für die Speicherung der taxonomischen Informationen die NCBI Taxonomie.

Für den Aufbau der GlycomeDB wurde eine Java Klassenbibliothek entwickelt, welche in der Lage ist, Kohlenhydratstrukturen zu laden, zu verarbeiten und zu speichern. Für den Import der Strukturen aus den verschiedenen Sequenzformaten, welche in den einzelnen Datenbanken ihre Anwendung finden, wurden diese Formate zunächst analysiert und die zugrundeliegenden Grammatiken extrahiert. Basierend auf diesen Grammatiken wurden dann Parser für den Import der Strukturen implementiert. Mit Hilfe von Wörterbüchern für Monosaccharidnamen wurden die importierten Strukturen danach in das Sequenzformat GlycoCT übersetzt. Auch für die taxonomischen Annoationen wurden Wörterbücher angelegt, welche eine Übersetzung der taxonomischen Zuordnungen in die NCBI Taxonomie erlauben.

Mit Hilfe der Klassenbibliothek, der Klassen zum Laden der Sequenzformate und den Wörterbüchern wurde das Java Programm GlycoUpdateDB erstellt, welches die Strukturen und taxonomischen Annotationen der sieben Datenbanken herunterlädt, vereinheitlicht und in der Datenbank GlycomeDB speichert. Abgesehen von der Pflege der Wörterbücher läuft das Programm automatisch und wird derzeit benutzt um die GlycomeDB wöchentlich mit den anderen Datenbanken abzugleichen. Während der Integration der Daten in der GlycomeDB werden Sequenzen, welche Fehler oder nicht übersetzbare Monosaccharide enthalten, vom Integrationsprozess ausgenommen und als Fehler in der GlycomeDB gespeichert. Basierend auf diesen Fehlerlisten wurden Berichte verfasst, die dann von den einzelnen Datenbankbetreibern benutzt wurden, um die Fehlerrate in ihren Datensätzen zu reduzieren.

Um die GlycomeDB und die gespeicherten Strukturen der wissenschaftlichen Gemeinschaft zur Verfügung zu stellen, werden die Datenbank und das Integrationsprogramm GlycoUpdateDB frei zum Herunterladen angeboten. Darüber hinaus wurde ein Webportal entwickelt, welches es erlaubt, online auf die GlycomeDB zuzugreifen. Das Webportal mit der URL www.glycome-db.org stellt zahlreiche struktur- und taxonomiebasierte Suchen bereit, mit welchen der Datenbestand der GlycomeDB und somit indirekt der Datenbestand der sieben integrierten Datenbanken durchsucht werden kann. Das neuartige Complex Query System erlaubt es darüber hinaus, mehrere verschiedene Suchanfragen miteinander zu kombinieren. Als Ergebnis wird eine Webseite mit allen Informationen über eine Kohlenhydratstruktur angegeben, in welcher auch die Einträge aus den sieben Datenbanken referenziert sind. Dies ermöglicht es dem Benutzer, auf diese Webseiten zu wechseln und weitere Informationen über die Struktur zu erhalten. Um auch eine automatisierte Suche nach Strukturen und Abfrage von Informationen zu ermöglichen, wurden zwei Webservice Schnittstellen implementiert. Über diese Schnittstellen ist es anderen Datenbanken und Programmen möglich, auf die Daten der GlycomeDB zuzugreifen und diese zu verwenden.

Als Ergebnis der Arbeit ist durch die Vereinheitlichung der Strukturen und taxonomischen Daten von sieben Datenbanken eine neue Ressource für Kohlenhydratstrukturen entstanden, welche den derzeit vollständigsten Index von Kohlenhydratstrukturen enthält. Durch das implementierte Webportal wird die Datenbank hauptsächlich als Suchmaschine für Kohlenhydratstrukturen eingesetzt. Darüber hinaus wurden die gespeicherten Strukturen auch für statistische Analysen verwendet.