Stephan Herget
Dr.sc.hum

Development of Databases, Bioinformatic Procedures and Analytics for Glycobiology

Geboren am 22.11.1974 in Usingen
Diplom der Fachrichtung Biochemie am 06.12.2002 an der Universität Tübingen

Promotionsfach: Biochemie
Doktormutter: Prof. Dr. Luise Krauth-Siegel

Die Arbeit spannt einen thematischen Bogen von der theoretischen Entwicklung eines umfassenden Sequenzformates für Kohlenhydratsequenzen zur Errichtung der weltweit ersten Metadatenbank für diese Substanzklasse. Die Anwendbarkeit elektronischer Datensammlungen für die Glykobiologie wird durch eine Studie zur Akzessibilität von Kohlenhydratsequenzen in der Festphasensynthese und eine vergleichende Studie zu bakteriellen und von Säugetierorganismen stammenden Kohlenhydraten untermauert. Zusätzlich wurde wichtige Infrastruktur für die Webseite der Metadatenbank während dieses Projektes impementiert.
Zu Beginn des Projektes wurden die existierenden Sequenzformate für die elektronische Speicherung von Kohlenhydratsequenzen evaluiert. Nach einer kritischen Bestandsaufnahme wurde erkennbar, daÿ keine der existierenden Lösungen alle vorhersehbaren Probleme der Strukturenkodierung konsistent beherrschen kann, so daß die Entwicklung eines neuen Formates in Angriff genommen wurde. Dieses neue Format mit dem Namen GlycoCT vereinigt alle Fähigkeiten der existierenden Formate in sich und wurde in zwei Varianten definiert. Herausragende Merkmale dieses Formates sind die maschinenlesbare Notation für Monosaccharide, der Formalismus der Gruppenersetzung bei Verknüpfung von Bausteinen und die Adressierbarkeit individueller Elemente durch numerische Identifikatoren. Da das Format durch einen Algorithmus sortiert werden kann, eignet es sich direkt als Primärschlüssel in Datenbankapplikationen.
Seit dem Ende der CarbBank, die die erste zentrale Datenbank in der Glykobiologie war, wurden verschiedene unabhängige Datensammlungnen für Kohlenhydratsequenzen entwickelt. Unglücklicherweise sind diese Datenbanken, im Gegensatz zu den etablierten Nukleotid- und Proteindatenbanken, nur unzureichend miteinander vernetzt und enthalten teils überlappende Sequenzinformation. Deswegen existieren mehrere Zugriffsportale für strukturelle Suchen. Das Ziel der GlycomeDB, einer neuen Datenbank, ist die Integration der Kohlenhydratsequenzen aller öffentlichen Datenbanken in eine einzige Metadatenbank. Bis dato wurden sieben Datenbanken integriert. Der Beitrag dieser Arbeit zur GlycomeDB liegt in der Entwicklung und Koordination der Datenaquisition, der Implementierung eines Übersetzungssystemes für die heterogene Nomenklatur der Monosaccharide und der Entwicklung strukturelbasierter Suchen und Klassiffzierungen. Die GlycomeDB wird wöchentlich mit den neuesten Sequenzen von sieben verschiedenen Quellen aktualisiert (siehe auch http://www.glycome-db.org).
Diese Arbeit wird durch zwei Studien abgerundet, die auf Datenbankanalysen und bioinformatischen Methoden basieren. Die erste Studie behandelt die Ableitung eines Satzes von Bausteinen für die festphasengestützte Oligosaccharidsynthese. Die zweite Studie beschäftigt sich mit einer systematischen Analyse der bakteriellen Saccharide im Vergleich zu Säugerdaten.