In: Hammwöhner, Rainer; Wolff, Christian; Womser-Hacker, Christa (Hg.): Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002), Regensburg, 8. – 11. Oktober 2002. Konstanz: UVK Verlagsgesellschaft mbH, 2002. S. 395 – 396

LINUCS - Eine Notation zur Unterstützung von Repräsentation und Retrieval spezieller chemischer Strukturen

Elke Lang[1], Andreas Bohne-Lang[2], Claus-Wilhelm von der Lieth[2], Alexander Loß[2]

[1]Fachhochschule Darmstadt
FB Informations- und Wissensmanagement
Max-Planck-Str. 2
64807 Dieburg
lang@iuw.fh-darmstadt.de

[2]Deutsches Krebsforschungszentrum
Zentrale Spektroskopie
AG Molecular Modeling
Im Neuenheimer Feld 280
69120 Heidelberg
a.bohne@dkfz.de

Zusammenfassung
Molekülstrukturorientierte Suche ist eine der wichtigsten Anforderungen an Substanzinformationssysteme. Da die Suche inzwischen meist von Anwendungsfachleuten ohne Kenntnis spezieller Retrievalsprachen durchgeführt wird, ist eine komfortable, intuitiv verständliche Benutzeroberfläche ein wesentlicher Akzeptanzfaktor. Beim Substanzinformationssystem SWEET-DB bildet ein Strukturgraph-Parser das Bindeglied zwischen der graphisch orientierten Eingabemaske für Strukturvorgaben und der (linearen) LINUCS-Graphnotation, die als Speicherformat für Zuckermolekül-Strukturen dient.

Abstract
Structure-oriented retrieval is an essential feature of substance information systems. Chemists most often perform standard retrieval projects on their own and therefore prefer user surfaces that allow input according to chemical notation rather than requiring knowledge of special retrieval languages. SWEET-DB, a glycosubstance information system, offers structure retrieval by using an input matrix that is fully conform to IUPAC nomenclature. A structure parser converts graphical 2D structure input into linear notation that is used for storage and graph matching.

1 Die LINUCS-Notation im Kontext der SWEET-DB

Die SWEET-DB [Loß2002] ist ein Substanzinformationssystem, das Molekülstrukturen einer bestimmten Substanzklasse enthält (Kohlenhydrate). Diese Spezialisierung erlaubt den Gebrauch entsprechender Notationen, die effi-

zienter sind als substanzklassenunabhängige Beschreibungsverfahren. Derartige Spezialnotationen haben sich z.B. für die Deskription und Suche von Proteinsequenzen eingebürgert [Berman2000]. Seit einiger Zeit werden Versuche unternommen, ein derartiges Verfahren auch für Kohlenhydrate zu entwickeln, wobei deren höhere strukturelle Komplexität, vor allem die Verzweigung, erhebliche Schwierigkeiten verursacht [Laine1994]. Die meisten Verfahren verlangen über die Standard-Nomenklatur [IUPAC1997] hinaus die Einhaltung weiterer Konventionen [Engelsen1996]. Die LINUCS-Notation [Lieth2001] bietet die Möglichkeit, Suchstrukturvorgaben mit einer Eingabematrix zu erstellen, in die wie gewohnt zweidimensionale (verzweigte) Strukturgraphen eingetragen werden können, die aus Monomertyp- und Bindungstyp-Elementen kombiniert werden. Der LINUCS-Parser wird zum einen verwendet, um bei der Aufnahme neuer Substanzdatensätze deren lineare Strukturnotation als Speicherformat zu erzeugen. Zum anderen wandelt er bei der Struktursuche die graphisch erstellte Strukturvorgabe in die lineare Notation um. Diese wird anschließend zum Graphvergleich benutzt.

Die LINUCS-Notation ist durch ihre lineare Form rechnergeeignet und ermöglicht eine schnelle Struktursuche; auch Ähnlichkeits- und Teilstruktursuche sind möglich. Bei der Aufnahme neuer Daten bietet LINUCS die Möglichkeit, neu aufgenommene Strukturen auf Plausibilität zu untersuchen.

Das LINUCS-Verfahren und die SWEET-DB sind unter http://www.dkfz-heidelberg.de/spec/ zugänglich, dort sind auch einführende Beispiele zu finden.

2 Literatur

[Engelsen1996] Engelsen SB, Cors S, Mackie W, Pérez S: A Molecular Builder for Carbohydrates: Application to Polysaccharides and Complex Carbohydrates. Biopolymers 39 (1996) 417-433

[IUPAC1997] IUPAC-IUBMB, Nomenclature of Carbohydrates. Carbohydrate Research 297 (1997) 1-92

[Laine1994] Laine RA: Calculation of all possible oligosaccaride isomers both branched and linear yield 1.05 x 10E12 structures of a reducing hexasaccharide. Glycobiology 4 (1994) 759-767

[Lieth2001] von der Lieth CW, Bohne-Lang A, Lang E, Förster T: LINUCS: Linear Notation for Unique description of Carbohydrate Sequences. Carbohydrate Research 336 (2001) 1-11

[Loß2002] Loß A, Bunsmann P, Bohne A, Loß A, Schwarzer E, Lang E, von der Lieth CW: SWEET-DB: an attempt to create annotated data collections for carbohydrates. Nucl. Acids. Res. 30 (2002) 405-408