Inaugural-Dissertation

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

Vorgelegt von

Apotheker Klaus Karl Lohmann

aus Wiedenbrück

Tag der mündlichen Prüfung: 2. Dezember 2003

Beiträge zur Glykobioinformatik

Entwicklung von Software-Werkzeugen für die Glykobiologie

Gutachter: Prof. Dr. Manfred Wießler
Prof. Dr. Nils Metzler-Nolte

Meinen Eltern

Herrn Prof. Dr. M. Wießler danke ich sehr herzlich für die Betreuung und seiner Bereitschaft die Arbeit vor der Fakultät für Biowissenschaften zu vertreten.

Herrn Prof. Dr. N. Metzler-Nolte danke ich sehr herzlich für die Übernahme des Korreferates.

Mein besonderer Dank gilt Herrn Dr. Claus-Wilhelm von der Lieth, für die großartige fachliche Unterstützung und seine unermüdliche Bereitschaft mir zu helfen.

Dem Deutschen Krebsforschungszentrum danke ich für die finanzielle und technische Unterstützung.

Bei der deutschen Forschungsgemeinschaft möchte ich mich für die finanzielle Unterstützung der vorliegenden Arbeit im Rahmen des Projektes ‚Glykodatenbanken' bedanken.

Ich versichere an Eides Statt,
dass ich die vorliegende Arbeit
selbständig und nur mit den
erlaubten Hilfsmitteln
durchgeführt habe.

1 Inhaltsverzeichnis

2 Abkürzungsverzeichnis

ASCII	American Standard Code for Information Interchange
ATP	Adenosintriphosphat
CDI	kontinuierliche Deionisation
CE	Kapillarelektrophorese
CGI	Common Gateway Inferface
CI	Chemische Ionisation
COX	Cyclooxygenase
EI	Elektronenstoß-Ionisation
ESI	Elektro-Spray Ionisation
EuAB	Europäisches Arzneibuch
FAB	Fast Atom Bombardment
GC	Gas Chromatography
GT	Glykosyltransferasen
GUI	Grafisches User Interface
HPLC	High Pressure Liquid Chromatography
IUPAC	International Union of Pure and Applied Chemistry
MALDI	Matrixunterstützte Laserdesorption/Ionisation
MS	Massenspektrometrie
MySQL	My Structured Query Language
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NLM	National Library of Medicine
PHP	PHP: Hypertext Preprocessor
QTOF	Quadrupol Time of Flight
RP	Reversed Phase
SQL	Structured Query Language
UV	Ultraviolet
VIS	Visuell

3 Einleitung

3.1 Allgemeines

Kohlenhydrate sind in der Natur weit verbreitet und haben eine große Bedeutung für die Biologie im Allgemeinen und die Pharmazie im Besonderen. So sind Kohlenhydrate in der Form von Traubenzucker (Fructose), Saccharose oder aber auch als Stärke ein guter Energielieferant und gleichzeitig ein Speicherstoff für Energie. Kohlenhydrate haben aber ebenfalls eine große Bedeutung für die Metabolisierung von Stoffen im menschlichen Körper. So werden z. B. schlecht wasserlösliche Stoffe im Körper glykosyliert, damit sie besser mit dem Urin ausgeschieden werden können[1]. In der Flora wird die Cellulose als Gerüstsubstanz beim Aufbau der Zellwände verwendet. Viele sekundäre Inhaltsstoffe werden mit Zuckerresiduen verknüpft. Unter anderem ist hier die Stoffgruppe der herzwirksamen Glykoside, die große Gruppe der Flavonoidglykoside[2, 3], aber auch die Cumaringlucoside von *Aesculum Hippocastanum L.* zu nennen[2, 3]. Eine der wichtigsten Aufgaben der Kohlenhydrate besteht aber in der Steuerung von Stoffwechselvorgängen und bei der Kommunikation in der Zell-Zellinteraktion[4].
Bei der Entschlüsselung des humanen Genoms hat sich herausgestellt, dass nicht mehr als Dreißig- bis Vierzigtausend verschiedene Proteine durch die DNA kodiert werden können[5]. Eine der überraschenden Erkenntnisse ist die im Vergleich mit anderen Organismen relativ geringe Anzahl von Genen. Es ist daher eine wichtige Herausforderung zu untersuchen, wie co- und posttranslationale Modifikationen die Eigenschaften und Funktionen von Proteinen beeinflussen und moderieren können. Es gilt daher als sicher, dass die Aktivität der Proteine, insbesondere der Enzyme, durch co- und posttranslationale Modifizierungen, hier vor allem durch Phosphorylierungen und Glykosylierungen, gesteuert wird. Deshalb sind in den letzten Jahren verstärkt die post- und cotranslationalen Veränderungen dieser Proteine in den Focus der Wissenschaft gelangt. Diese Veränderungen können aus einer Vielzahl von chemischen Stoffen bestehen, die kovalent an das jeweilige Protein gebunden sind[6].
Glykosylierung ist eine der am häufigsten vorkommenden Modifikationen, die strukturell sehr vielfältiger Art sein können. Die physiologischen Veränderungen, die ein Protein- durch die Glykosylierung erfährt, werden in intrinsische und extrinsische Faktoren untergliedert. Dabei versteht man unter intrinsischen Faktoren solche, die als strukturelle Komponenten dienen und z.B. zur Stabilisierung einer Proteinkonformation beitragen oder solche, die die physikochemischen Eigenschaften wie z.B. Löslichkeit beeinflussen. Unter den extrinsischen Funktionen fasst man alle biologischen Erkennungsprozesse zusammen an denen Glykane beteiligt sind.

Intrinsische Funktionen	Extrinsische Funktionen
Struktureller Bestandteil von Zellwänden und der extrazellulären Matrix	Modulierung der Zelladhäsion durch Zell-/ Zell und Zell-/ Matrix Interaktionen
Beeinflussung der Löslichkeit und Stabilität eines Proteins	Beeinflussung und Modulierung der intra- und extrazellularen Signale
	Lotsenfunktion für die inner- und extrazellularen Glykokonjugate

Tabelle 1: In- und extrinsische Funktionen von Glykosylierungen

Die Schlüsselrolle dieser Modifikationen in der Biologie besteht in der Aktivierung bzw. Deaktivierung von Proteinen und bei der Erkennung von Strukturen, z .B. auf der Zellmembran[7].
Die spezifischen Glykanstrukturen, die sich an einer bestimmten Glykosylierungsstelle des Proteins ausbilden, werden nur indirekt durch das Genom bestimmt. Diese Arbeit verrichten Enzyme, die unter dem Begriff Glykosyltransferasen (GT) zusammengefasst werden. Sie bauen in einer sehr spezifischen Synthese N- und O-Glykane auf. Die Informationen zum Aufbau dieser Glykane ist in den GT kodiert, die einer Zelle zur Verfügung stehen[8]. So gibt es mehrere Hundert verschiedene GT im menschlichen Körper, deren einzige Aufgabe es ist, Sequenzen von Zuckern zu bilden. Eine Datenbank mit allen kohlenhydrataktiven Enzymen(CAZy)[1] steht im Internet zur Verfügung und bietet einen guten Überblick über die Vielzahl der unterschiedlichen Enzyme, die am Ab- und Aufbau von Kohlenhydraten beteiligt sind[9, 10]. Zu den kohlenhydrataktiven Enzymen gehören die Abteilungen der Glykosidasen, Transglykosidasen, Glykosyltransferasen, Polysaccharid-Lyasen und Kohlenhydrat-Esterasen. Auf eine der Glykosidasen, eine Neuramidase, wird in einem der nächsten Abschnitte noch genauer eingegangen.

3.2 Co- und posttranslationale Modifikationen von Proteinen

Nachdem die mRNA den Zellkern verlassen und an den Ribosomen die Translation begonnen hat, kann es während dieser Translation (cotranslational) und auch danach (posttranslational) zu einer kovalenten Veränderung einer einzelnen Aminosäure des Proteins kommen. Es gibt die unterschiedlichsten Arten der Modifikationen. Diese gehen von relativ einfachen Molekülen, wie einer Phosphatgruppe oder einer Methylierung bis zu sehr komplexen Glykosylierungen, die aus bis zu 20 verschiedenen Residuen bestehen können.

[1] http://afmb.cnrs-mrs.fr/CAZY/

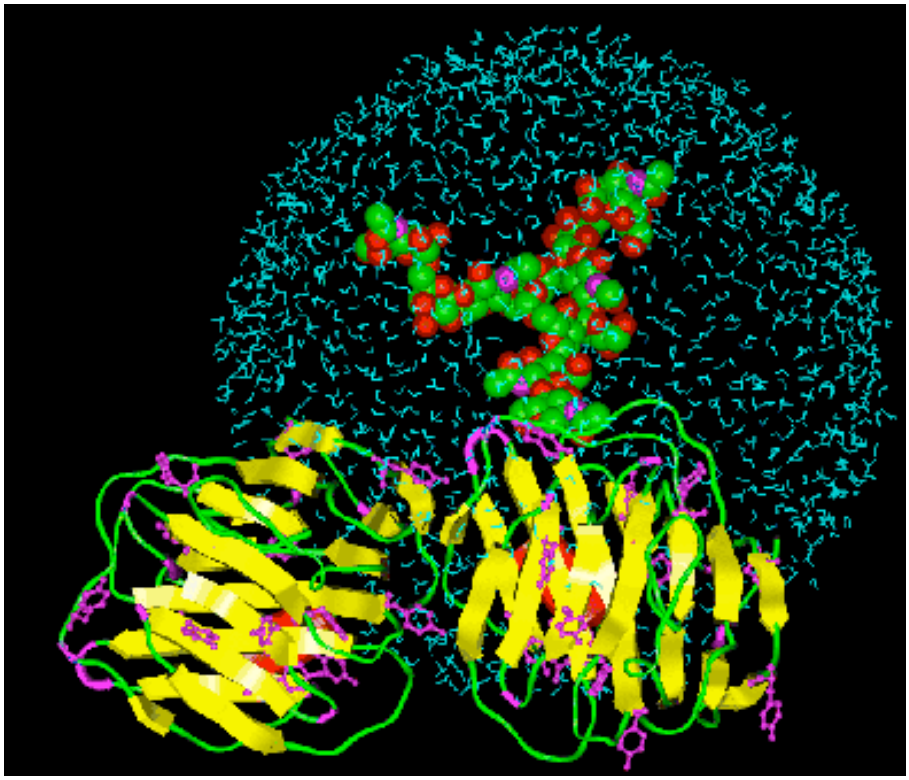


Abbildung 1: 3D-Darstellung eines glykosilierten Proteins (Die Glykosylierung ist solvatisiert)

3.2.1 Phosphorylierung

Etwa 30-50% Prozent aller zellulären Proteine sind phosphoryliert. Dazu wird kovalent und reversibel eine Phospat-Gruppe an eine Aminosäure des Proteinstrangs gebunden. Diese Phosphorylierung wird häufig als molekularer Schalter zur Veränderung von Proteinen benutzt. Ein Beispiel dafür ist die Pyruvatdehydrogenase, die durch eine reversible Phosphorylierung reguliert wird[11]. Mittels Proteinkinasen ist es möglich, eine Phosphatgruppe unter Verwendung von ATP auf eine Aminosäure zu übertragen. Diese Bindung kann durch Phosphatasen wieder gelöst werden[12]. Auf Grund der Regelfunktion dieser Phosphorylierungen sind diese Enzyme häufig Ziele für Medikamente, unter anderem auch für die Hemmung der Mitogen-aktivierten Proteinkinase [13, 14]. Der Einsatz dieser Substanzen findet im Bereich des Darmkrebses[15] und der entzündlichen Erkrankungen[16] statt.

3.2.2 Glykosylierungen

Die Zuckerstrukturen sind entweder über ein Stickstoffatom der Aminosäure Asparagin (N-Glykane) oder über das Sauerstoffatom von Serin oder Threonin (O-Glykane) mit dem Protein verknüpft. Während für die N-Glykane eine Erkennungssequenz auf Proteinebene existiert (Asn-X-Ser/Thr, wobei X ungleich Prolin ist), kann praktisch jedes auf der Proteinoberfläche exponierte Serin oder Threonin glykosyliert werden. Der Aufbau der Glykane und auch die Stellen der Glykosylierungen sind sehr spezifisch, da nur eine endliche Menge unterschiedlicher Glykosyltransferasen zur Verfügung stehen. Obwohl die gleichen Glykosylierungs-werkzeuge allen Kopien eines Proteins zur Verfügung stehen, die einen bestimmten Stoffwechselweg einer Zelle durchlaufen, sind viele Proteine nicht mit strukturell einheitlichen Glykanen versehen, sondern zeigen ein Muster, das für eine jede Glykosylierungsstelle charakteristisch ist. Aus diesem Grunde ist das

Glykosylierungsmuster ein sehr empfindlicher Marker für Veränderungen in den Zellen. So können mittels dieser Glykosylierungen folgende Parameter unterschieden werden.

1. Spezies[17]
2. Gewebe[18]
3. Protein[19]

Die Analytik dieser Zuckerstrukturen von Proteinen ist daher von großem Interesse für die Medizin und Pharmazie, da z.B. ein Unterschied im Glykosylierungsmuster auf eine mögliche Krankheit oder einen Enzymdefekt hinweisen kann. So hat sich bei der Untersuchung der Zuckerstrukturen von Prionen, die für den Ausbruch der BSE-Krankheit verantwortlich gemacht werden, herausgestellt, dass ein Unterschied in der Glykosylierung der Prionen zum Ausbruch der Krankheit führen kann[20]. Hat man zudem noch die Möglichkeit den Stoffwechselweg bei der Bildung der Glykosylierung zurückzuverfolgen, hat man zugleich ein mögliches Target für eine Therapieform gefunden.

3.2.2.1 N-Glykane

Bei den Glykosylierungen bilden die N-Glykane die größte Gruppe. So sind sehr wahrscheinlich bis zu siebzig Prozent aller Proteine N-glykosyliert[21] und gehören auch zur bestuntersuchten Gruppe der Modifikationen. Auf Grund der unterschiedlichen glykosidischen Verknüpfungsmöglichkeiten bei den einzelnen Zuckerresiduen und auch der großen Anzahl möglicher Monosaccharidbausteine ist eine sehr große Anzahl unterschiedlicher Verbindungen dieser Substanzklasse denkbar, die sich aus der Kombination der vorhandenen kohlenhydrataktiven Enzyme bilden lassen. Allein in der *SWEET-DB* befinden sich 7983[1] verschiedene Einträge, die als N-Glykane charakterisiert sind. Bei allen diesen Verbindungen ist aber ein Corebereich identisch. Dabei handelt es sich um ein Pentasaccharid, das aus drei β-D-Mannosen und aus zwei β-D-N-Acetyl-Glucosaminen besteht. Die hier verwendete Nomenklatur entspricht nicht zu 100% der IUPAC-Nomenklatur[22, 23]. Sie ist aber für die Datenverarbeitung effizienter. Eine genaue Beschreibung befindet sich im Abschnitt *GLYCO-FRAGMENT*.

```
a-D-Manp-(1-6)+
              |
         a-D-Manp-(1-4)-a-D-GlcNac-(1-4)-a-D-GlcNac
              |
a-D-Manp-(1-3)+
```

Abbildung 2: Core-Struktur eines N-Glykans

Dieser Corebereich kann an dem endständigen GlcNAc-Residuum mit einer α-L-Fucose mittels einer 1-6-glykosidischen Bindung erweitert werden.

[1] Stand September 2003

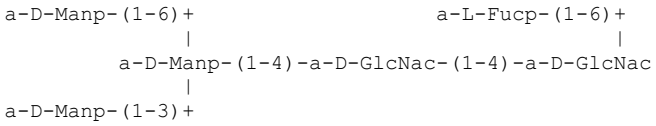


Abbildung 3: Core-Struktur eines N-Glykans mit proximaler Fucose

Diese Verbindungen werden dann als core-fokusyliert bezeichnet. An dem zweiten N-Acetyl-Glucosamin Residuum kann sich 1-4 glykosidisch verknüpft ein weiteres N-Acetyl-Glucosamin befinden. Glykane, die ein derartiges Residuum besitzen, werden als „bisected“ bezeichnet.

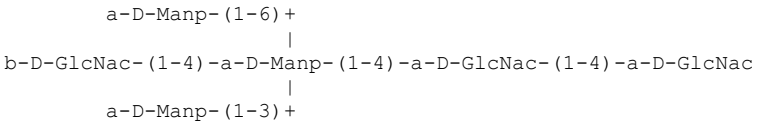


Abbildung 4: Core-Struktur eines N-Glykans mit bisecting GlcNac-Residuum

Sämtliche Glykane können sowohl mit einer proximalen Fucose als auch mit einem bisecting N-Acetyl-Glucosamin Residuum vorkommen. Diese Core-Bereiche sind nun über das C1-Atom des □-D-N-Acetyl-Glucosamin Residuums mit dem Amidstickstoff eines L-Asparigin der Peptidsequenz verknüpft. Innerhalb der Peptidsequenz kann die N-Glykosylierung nur an diesen beiden Aminosäurefolgen (wobei Xxx ungleich Prolin ist) erfolgen:

Asn-Xxx-Ser
Asn-Xxx-Thr

An diese Kern-Verbindungen können durch Glykosyltransferasen sehr spezifisch weitere Residuen glykosidisch gebunden werden. Wie oben beschrieben sind diese Transferasen sehr spezifisch und bilden den limitierenden Faktor bei der Synthese der N-Glykane Die daraus resultierenden Zuckerstrukturen sind sehr vielfältig. Prinzipiell lassen sich diese Verbindungen jedoch grob in drei Strukturklassen aufteilen.

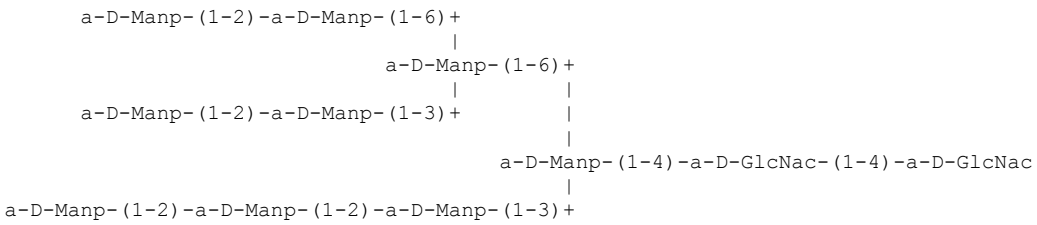


Abbildung 5: High-Mannose Typ

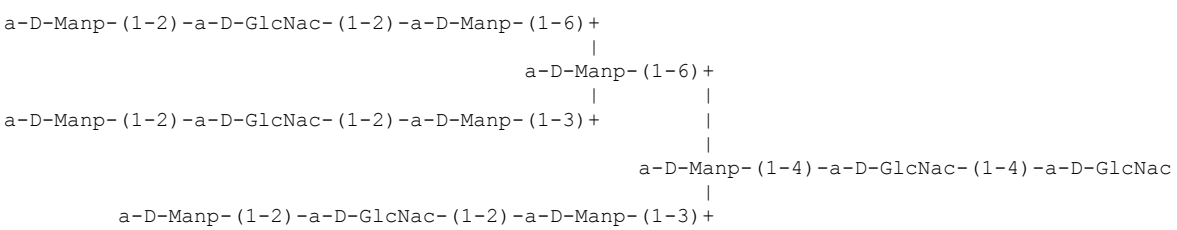


Abbildung 6: Complexed Typ

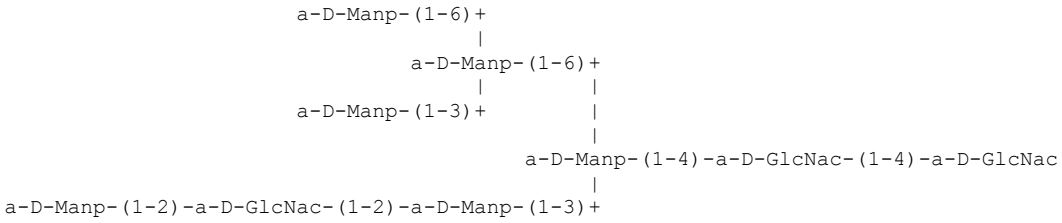


Abbildung 7: Hybrid Typ

Zur weiteren Unterteilung der N-Glykane wird auch sehr häufig die Anzahl der Antennen herangezogen. Dabei wird die Anzahl der Residuen genommen, die glykosidisch mit den beiden endständigen Mannosen der Core-Struktur verbunden sind. In der Natur kommen sehr häufig bi-, tri-, und tetraantennäre Glykane vor. In Einzelfällen werden aber auch mono- und pentaantennäre Glykane beobachtet[24-26].

3.2.2.2 O-Glykane

Die O-Glykane besitzen ebenso wie die N-Glykane eine Anzahl von Core-Strukturen, von denen ausgehend dann die weiteren O-Glykane abgeleitet werden. Die O-Glykane werden über ein Sauerstoff-Atom kovalent an die Aminosäuren Serin oder Threonin mit dem Protein verbunden[27]. Als erstes wird dazu ein N-Acetyl-Galactosamin-Residuum enzymatisch mit dem Protein verbunden. An dieses N-Acetyl-Galactosamin-Residuum werden dann durch Glycosyltransferasen weitere Zuckerbausteine glykosidisch gebunden.

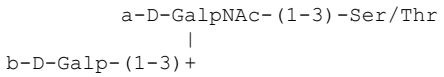


Abbildung 8: Thomson-Friedenreich-Antigen/Core 1

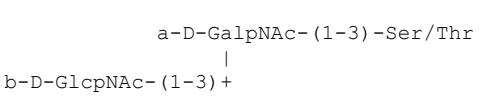


Abbildung 9: Core 2

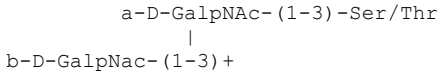


Abbildung 10: Core 3

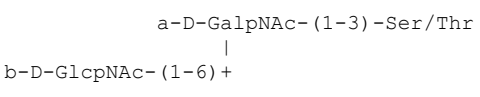


Abbildung 11: Core 4

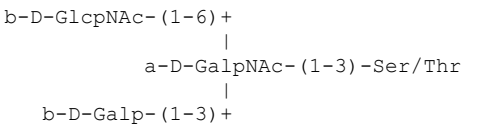


Abbildung 12: Core 5

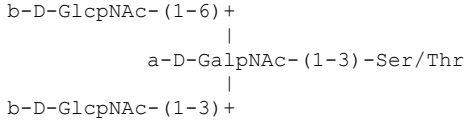


Abbildung 13: Core 6

Die biologische Funktion der O-Glykane ist noch Gegenstand der Forschung. Es wird aber vermutet, dass sie zur Stabilisierung der Tertiärstruktur des Proteins benötigt werden. Auch bei den O-Glykanen existiert eine große strukturelle Vielfalt der experimentell nachgewiesenen Strukturen. Es lassen sich aber keine Klassen wie bei den N-Glykanen definieren.

3.2.2.3 Glykolipide

Anders als bei den Glykoproteinen, bei denen Polysaccharidketten an Proteine gebunden sind, werden bei den Glykolipiden Zuckerstrukturen an Lipide gebunden. Dadurch erhalten diese Moleküle einen amphiphilen Charakter. So besitzen diese Strukturen einen hydrophilen Saccharidbereich und einen lipophilen Lipidanteil. Der Lipidanteil dient dazu, die Glykanstrukturen in der Zellmembran zu verankern.
Die wichtigsten Klassen der Glykolipide sind zum einen die Sphingolipide und zum anderen die Inositole. Bei den Sphingolipiden handelt es sich um Zuckerresiduen, die glykosidisch mit einem Ceramid verbunden sind.

3.3 Bedeutung der Glykosylierungen für die Pharmazie und Medizin

In der Pharmazie spielen zwei Arten von Glykosylierungen eine entscheidende Rolle: Als erstes sind hier die Arzneistoffe zu nennen, die in ihrer Wirkform einen Zuckeranteil aufweisen. Häufig sind dies sekundäre Inhaltsstoffe aus dem Bereich der pharmazeutischen Biologie. Die pharmazeutisch wichtigste Gruppe bildet die der herzwirksamen Glykoside. Als wichtigste Vertreter dieser Arzneistoffklasse sind hier das Digoxin(Lanicor®) und das Digitoxin(Digimerck®) zu nennen:

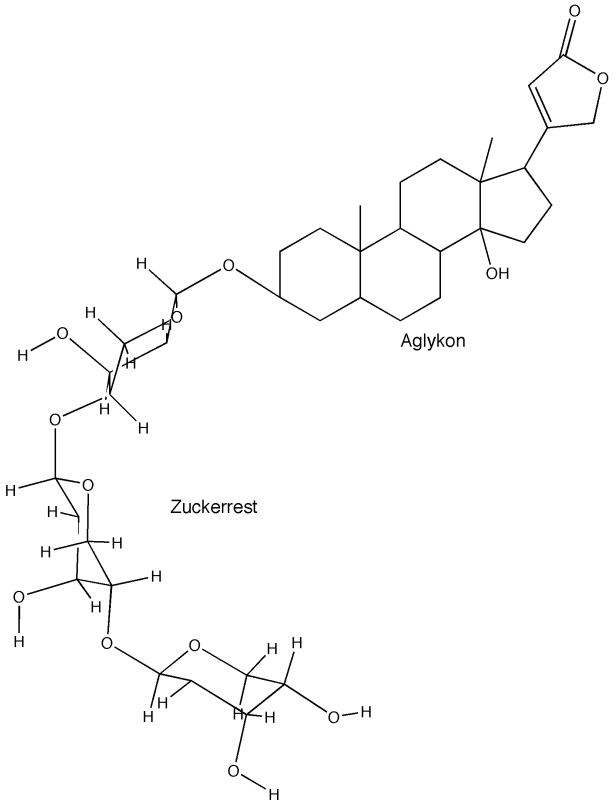


Abbildung 14: Strukturformel von Digitoxin

Alle herzwirksamen Glykoside bestehen aus einem Aglykon und einer Zuckerkette, deren Zusammensetzung unterschiedlich ist. Die Zuckerkette hat zwar keinen direkten Einfluss auf die arzneiliche Wirkung, je nach Art und Anzahl der Residuen ändert sich aber die Halbwertszeit der Substanzen im Serum. Es gibt noch eine ganze Reihe weiterer glykosidischer Verbindungen, die als Arzneistoffe eine Rolle spielen, es sind dies unter anderem die Saponine und Aminoglykoside[28].

3.3.1 Kohlenhydrat-aktive Enzyme als Target für Medikamente

Als mögliche Targets für die Medikamentenentwicklung sind die Glykosyltransferasen und Glykosidasen wesentlich interessanter, da durch die Blockierung der katalysierten Reaktionen dieser Enzyme sehr spezifisch in den Stoffwechsel eingegriffen werden kann. So gibt es in der Regel für den Aufbau einer bestimmten glykosidischen Bindung aus zwei Substraten jeweils nur ein Enzym in der Zelle, das diesen Schritt vornehmen kann[7]. Ein aktuelles Beispiel für diese rationale Vorgehensweise ist die Entwicklung neuartiger antiviraler Medikamente gegen Influenza. Die beiden Arzneistoffe Zanamivir und Oseltamivir hemmen das Enzym Neuraminidase, so dass die Ablösung des Virus, und damit seine Vermehrung von der Zelloberfläche der Wirtzelle unterbleibt. Die neuen Arzneistoffe ähneln dem natürlichen Substrat des Enzyms der Sialinsäure, die als terminale Residuen von Gangliosiden in der Zellmembran der Wirtzelle verankert sind. Der Guanidinorest im Zanamivir, der anstelle einer Hydroxylgruppe eingebaut ist, sorgt für die Verdrängung der Sialinsäure und eine feste Bindung des Arzneistoffs im aktiven Zentrum.

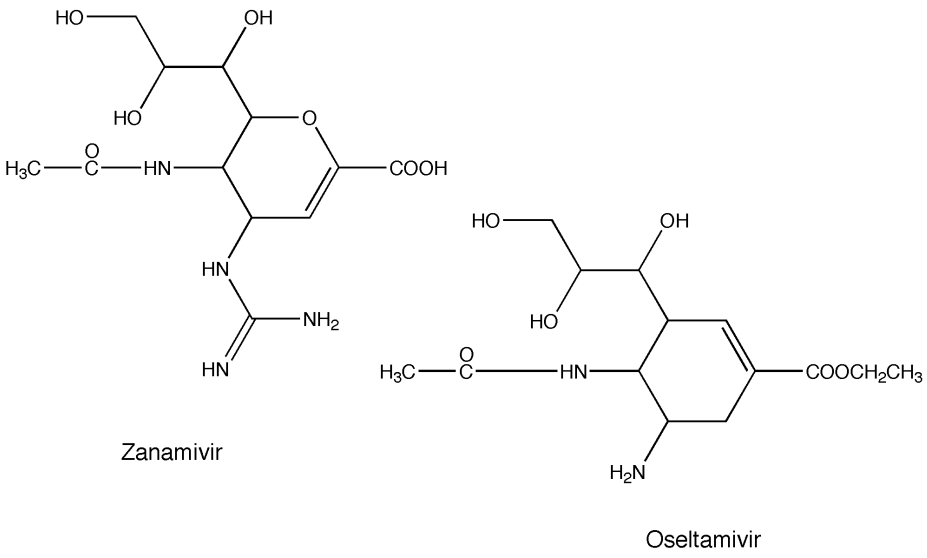


Abbildung 15: Strukturformeln von Zanamivir und Oseltamivir

Ein großer Nachteil von Zanamivir (Relenza®)[29] besteht in der schlechten Bioverfügbarkeit dieser Substanz, so muss die Substanz über einen Pulverinhalator appliziert werden. Außerdem besitzt diese Substanz einen sehr hohen Anteil an Nebenwirkungen, was sehr wahrscheinlich durch das ubiquitäre Vorkommen der Glykosyltransferasen erklärt werden kann. Oseltamivir(Tamiflu®)[30] ist die zweite Substanz, die eine Zulassung für die Behandlung der Influenza A und B erhalten hat. Oseltamivir ist deutlich lipophiler als Zanamivir und kann daher auch peroral gegeben werden.
Beide Substanzen sind mittels Molecular Modelling in das aktive Zentrum gebaut worden und entsprechen in ihrer Elektronenstruktur genau dem Substrat Sialinsäure[31-33]. Im Gegensatz dazu können Zanamivir und Oseltamivir, deren Bindungskonstanten größer sind als die des Substrates Sialinsäure, das aktive Zentrum aber nicht mehr verlassen, so dass es zu einer nicht reversiblen Inhibierung der Neuraminidase kommt.
Dieses Beispiel zeigt, wie intensiv die kohlenhydrataktiven Enzyme an der Stoffwechselphysiologie des Organismus beteiligt sind, und wie wichtig es ist, diese Vorgänge auf der atomaren Ebene zu verstehen.

3.4 Analytische Methoden

Weltweit wird in verschiedenen Laboratorien intensiv daran gearbeitet, geeignete analytische Strategien zu entwickeln, die eine schnelle Kartierung des Glykoms eines Proteins ermöglichen. Das Glykom beschreibt die Gesamtheit aller Zuckerstrukturen für eine Spezies. Ähnlich wie im Bereich der Proteomik ist die Massenspektrometrie (MS) besonders gut geeignet für eine Analyse der Glykome. Generell ist die Bestimmung der Struktur von Glykanen aufgrund der unterschiedlichen Verknüpfungsmöglichkeiten der monomeren Einheiten wesentlich aufwendiger als die Analyse von DNA- und Proteinsequenzen. Andererseits ist mittels verschiedener NMR-Techniken eine vollständige Strukturbestimmung, wie die Art der glykosidischen Verknüpfung und die Konformation zugänglich. Allerdings liegt die benötigte Menge an reiner Substanz um mehrere Größenordnungen höher als bei den MS-Verfahren, so dass NMR-Techniken weniger gut geeignet sind für Hochdurchsatzverfahren.

3.4.1 Methoden zur Auftrennung eines Kohlenhydratgemisches

In der Regel liegen Kohlenhydrate nicht als Reinsubstanz vor, da es sich entweder um die Menge aller Glykosylierungen eines Proteins oder sogar einer Art handelt, die mittels chemischer[34, 35] oder enzymatischer Spaltung[36, 37] gewonnen worden sind. Für eine eingehende Untersuchung der einzelnen Kohlenhydrate ist es daher zwingend nötig, diese zu trennen. Dieses kann zum einen durch die Verwendung von Lektinen[38] oder durch Gelelektrophorese[37] geschehen, aber auch durch die Trennung mittels chromatographischer Verfahren wie HPLC[34, 36], GC[39] oder CE[40].

3.4.2 Massenspektrometrische Untersuchungsmethoden in der Biochemie

Die MS hat einen sehr hohen Stellenwert für den exakten Nachweis von biologischen Makromolekülen in physiologischen Zusammenhängen. Ein großer Vorteil der MS besteht in ihrer Empfindlichkeit, so dass sogar geringe physiologische Probenmengen bis hin in den femtomolaren Bereich[6] analysiert werden können, wie sie bei der Auftrennung von biologischen Proben üblicherweise anfallen. Außerdem ist die MS sehr vielseitig einsetzbar. So kann mit ihrer Hilfe nicht nur die Masse eines Proteins bestimmt werden. In Kombination mit Datenbanksuchen kann auch die komplette Sequenz eines Proteins[41], vorausgesetzt das Protein ist schon bekannt, oder eines Glykans[42] ermittelt werden. Die Aufbereitung der Proben und das Messprotokoll können auch so ausgerichtet werden, dass die Position der modifizierten Aminosäure bestimmt werden kann[43]. Unter anderem dienen sie der Ermittlung von Phosphorylierungsstellen[44, 45] und Glykosylierungsstellen [43, 46, 47].

3.4.3 Methoden zur Proteinsequenzanalyse

Die Bestimmung der Sequenz eines Proteins ist schon sehr lange möglich und im Focus der Wissenschaft. So konnte schon im Jahre 1926 mit Thiocyanat-Reagenz nach einem Verfahren von Schalck und Kumpf[48] die Sequenz bestimmt werden. Dieses Verfahren ist im Jahre 1991 von Inglis[49] verbessert worden, und die Methode steht inzwischen auch vollautomatisch zur Verfügung. Dabei wird mit Hilfe des Thiocyanat-Reagenzes durch Kupplung das C-Terminale Ende des Peptides aktiviert und anschließend abgespalten. Dieses abgespaltene Ende kann nun mittels RP-HPLC analysiert werden. Leider sind für diese Methode immer noch Substanzmengen von mehr als einem Nanogramm nötig.
Von der N-Terminalen Seite besteht die Möglichkeit des Edman-Abbaus[50]. Dabei wird durch Kupplung von Phenylisothiocyanat das N-Terminale Ende des Peptids aktiviert und anschließend abgespalten. Auch hier wird die entstandene Phenylthiohydantoin-Aminosäure mittels HPLC identifiziert.
Diese Verfahren sind natürlich sehr aufwendig, und es wird eine große Menge des zu analysierenden Peptides benötigt. In den letzten Jahren hat sich als Alternative die MS zur Sequenzierung von Peptiden etabliert, die bis zu maximal 20 Aminosäuren lang sein dürfen. Damit ein Protein sequenziert werden kann, müssen diese erst enzymatisch gespalten werden, um Bruchstücke in der richtigen Länge zu erhalten. Dazu stehen verschiedene Enzyme zur Verfügung, die jeweils charakteristisch an bestimmten Stellen den Proteinstrang aufspalten.

Enzym	Spaltung bei	Durchschnittliche Fragmentlänge
Chymotrypsin	Leucin Phenylalanin Tryptophan Tyrosin	6
Trypsin	Lysin Arginin	9
Endoproteinase GluC	Glutamat	15
Endoproteinase LysC	Lysin	16
Endoproteinase AspC	Aspartat (C-Terminal)	18
Endoproteinase AspN	Aspartat (N-Terminal)	18

Tabelle 2: Spezifische Spaltung von Enzymen[6]

Bei der Sequenzierung von Proteinen wird vor allem Trypsin als Enzym verwendet, da es in der Lage ist, Proteine sehr spezifisch nach den Aminosäuren Lysin und Arginin zu spalten. Die erhaltenen Bruchstücke besitzen eine für die weitere Analytik ideale Länge.

3.4.3.1 Tryptischer Verdau eines Proteins

Die Abbildung 16 zeigt beispielhaft den Verdau des Enzyms ‚Dynamin like protein' von *Dictyostelium discoideum* durch die Protease Trypsin. Dabei sind die enzymatisch gespaltenen Bindungen rot hervorgehoben. Die dabei erhaltenen Fragmente können nun massenspektrometrisch analysiert werden.

MDQLIPVINKLQDVFNTLGSDPLDLPQIVVVGSQSSGKSSVLENIVGRDFLPRGSGIVTRRP
LILQLTHLPIADDGSQTQEWGEFLHKPNDMFYDFSEIREEIIRDTDRMTGKNKGISAQPINL
KIYSPHVVNLTLVDLPGITKVPVGDQPTDIEQQIRRMVMAYIKKQNAIIVAVTPANTDLANS
DALQLAKEVDPEGKRTIGVITKLDLMDK**GTDAMEVLTGR**VIPLTLGFIGVINRSQEDIIAKK
SIRESLKSEILYFKNHPIYKSIANRSGTAYLSKTLNKLLMFHIRDTLPDLKVKVSKMLSDVQ
GELSTYGDPLYDTKNSQGALLLQIITIFSSNFKDAIDGKLTDLSNNELYGGARISYIFNEIY
SHCVNNIDPLEGISLNDIRTTMRNATGPRAALFIPEISFELLVKKQVVRLEEPSAQCVEYVY
DELQRIVSQLEAKELSRFINLKARVIEVVNNLLQKHKVPTKTMIEHLIKIETAFINTSHPDF
VGGEGIFESLYKKQQLQQQNHLQQLQDQYQQQQQQQQQQQQQNGINNNQKGDNGNMNVNQQN
MNQQNMNQQNQSTNPFLQQQQQGQNKYPGGPPAQQQPNQQPNQLNKGPQNMPPNQSKPSSIP
QNGPNNNNNNNNNNNRQDHQQGSFFSSFFRASPDPSLGQYGGANNSNNSNNPTSPINSSSNS
GNNYNTFGGQQSSSSSSQQLQQSSQSQYKTSYNNNNNSSSNNSSYNRYQDDFYGRGDKLNQV
PSIIKAPDDLTSKEKFETELIRELLISYFNIVKKNVKDSVPKSIMHFLVNQSKEHIQNELVA
ALYKEELFDELLEESPQISSKRKSCKAMIEILRKANEIINEIRDFRN

Abbildung 16: Schnittstellen beim tryptischen Verdau der Dynamin like protein

Daraus ergeben sich nun 85 Fragmente und auch 85 MS/MS-Spektren. In der Abbildung 17 wird anhand des Bruchstückes mit der Sequenz **GTDAMEVLTGR** die übliche Nomenklatur zur Bezeichnung der Ionen vom N-terminalen und C-terminalen Ende angegeben. Bei einer geeigneten Wahl der Messbedingungen lassen sich diese Bruchstücke bevorzugt in den erhaltenen MS/MS-Spektren wieder finden, und sie gestatten so eine einfache Sequenzierung des Peptids.

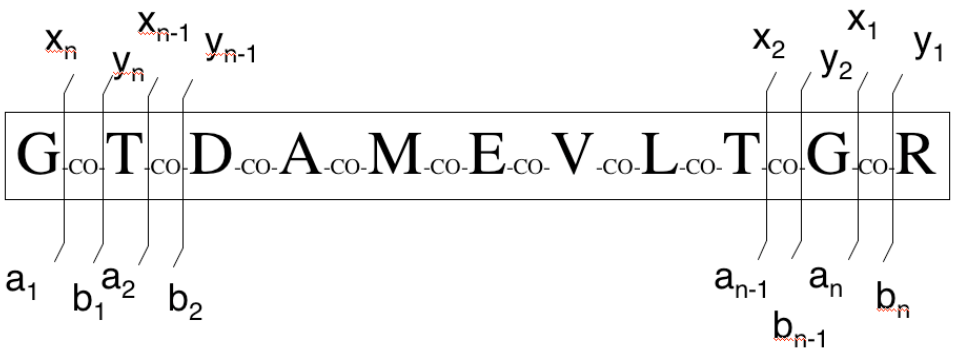


Abbildung 17: Bezeichnung der Fragmente von Peptiden, die bei der massenspektrometrischen Untersuchung auftreten. Dargestellt an einer durch Verdau mit Trypsin entstandenen Sequenz eines ‚Dynamin like protein'

Diese Spektren enthalten neben den reinen Peptid-Bruchstücken auch Bruchstücke an denen sich noch kovalent gebunden, die Modifikationen befinden.

3.4.3.2 Positionsbestimmung von co- und posttranslationalen Modifikationen

Einen sehr hohen Stellenwert in der Biochemie hat die Fragestellung, ob ein Protein eine post- oder cotranslationale Modifikation aufweist, und welche Aminosäure in dem Proteinstrang von dieser Veränderung betroffen ist. Denn zur Bestimmung eines Stoffwechselzustandes einer Zelle ist es von großer Bedeutung, ob und wo ein Protein nachträglich verändert worden ist.
Dabei wird wie folgt vorgegangen: Als erstes wird das Protein mittels eines enzymatischen Verdaus, wie oben beschrieben, in kleine Peptidstränge mit einer Masse in der Regel von weniger als 1500Da zerlegt. Diese Teilsequenzen des Proteins werden dann mittels MS/MS-Untersuchungen in die a-, b-, x- und y-Fragmente aufgesplittet. Ist bei einem dieser Fragmente nun eine der Aminosäuren verändert, so ergeben sich daraus Fragment-Ionen, die um einen charakteristischen Massenwert erhöht sind, wie in Tabelle 3 gezeigt.

Massendifferenz	Modifikation
80,0000	Phosphat-Gruppe
146,0600	d-Hexose
162,1424	Hexose
204,1876	N-Acetyl-Hexamin
14,0628	Methylierung

Tabelle 3: Typische Massendifferenzen und die entsprechende Modifikation

Um nun die genaue Position bestimmen zu können, sollten Serien dieser Fragment-Ionen in den auszuwertenden Massenspektren vorhanden sein. Für eine automatische Erkennung dieser Modifikationen stehen zurzeit noch keine geeigneten Softwareanwendungen zur Verfügung. Im Rahmen dieser Arbeit wird ein Algorithmus beschrieben, der in der Lage ist, diese Lücke zu schließen. Zum Test der Richtigkeit des Algorithmus entstand das Programm *FINDYSERIES*.

3.4.4 Bestimmung des Molekulargewichtes von Proteinen

Dabei wird das Protein zusammen mit einer in einem Lösungsmittel gelösten kleinmolekularen Gerüstsubstanz, wie z.B. Nicotinsäure, in eine Form gegeben. Nach Verdampfen des Lösungsmittels erhält man eine Matrix, in die eingebettet, sich

das Protein befindet. Durch kurzes Einwirkenlassen eines Laserstrahls verdampfen geringe Mengen des Proteins, und dieses kann dann ionisiert werden. Dabei kommt es zu keiner Zerstörung des Proteins, sondern man erhält das gesamte Protein, das allerdings unterschiedlich geladen sein kann[6].

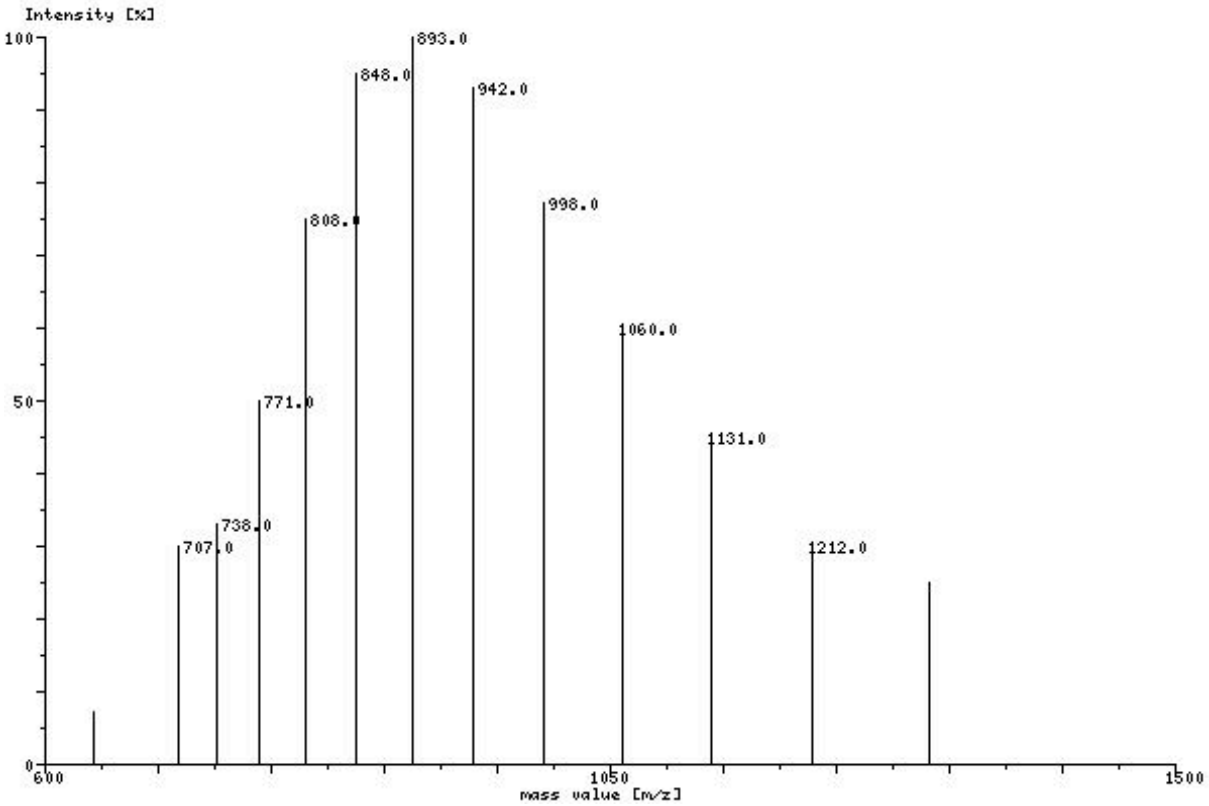


Abbildung 18: Beispielspektrum zur Bestimmung der Proteinmasse

Aus den Differenzen dieser Massen kann man dann Rückschlüsse auf die Masse des Proteins ziehen. Dabei wird wie folgt vorgegangen: Unter der Annahme, dass sich zwei benachbarte Peaks nur in einer Ladung unterscheiden und die Ladungen durch das Anlagern von Protonen entstehen, ergibt sich für einen Peak p_1 mit der Ladung z_1 die folgende Formel

$$p_1 z_1 = M_r + M_a z_1 = M_r + 1{,}0079\, z_1$$

mit M_a = Masse des angehängten H^+
Für den benachbarten Peak p_2 ergibt sich nun

$$p_2 (z_1 - 1) = M_r + 1{,}0079 (z_1 - 1)$$

Setzt man für die beiden Peaks die p_1= 1060,0 und p_2 = 998,0, so ergibt sich daraus eine Masse für das Protein von 18015Da.

3.5 Massenspektrometrische Untersuchung von Kohlenhydraten

Ähnlich wie bei der Identifizierung von Proteinen in der Proteomik zeichnet sich ab, dass auch die MS bei der Untersuchung eines Glykoms die analytische Methode der Wahl zur schnellen Bestimmung der Gesamtheit der Glykane ist. Der automatische

Vergleich der Massenspektren von enzymatisch verdauten Proteinen mit den theoretisch berechneten Massenspektren der entsprechenden Proteinfragmente erlaubt in vielen Fällen eine eindeutige Identifizierung der untersuchten Proteine. Bisher sind keine vergleichbaren Verfahren und Algorithmen für den Glykomik-Bereich verfügbar.

3.5.1 Derivatisierung von Kohlenhydraten

Zur Steigerung der Sensitivität und zur besseren Aufklärung der Art der glykosidischen Verknüpfung zwischen zwei Residuen wird häufig eine Derivatisierung der Ausgangsstruktur vorgenommen. Im ersten Fall wird häufig das Residuum am reduzierenden Ende verändert. Im zweiten Fall wird in der Regel eine Permethylierung oder Peracetylierung der gesamten Verbindung[35] durchgeführt. Um die an der glykosidischen Bindung beteiligten Atome der Residuen zu ermitteln, müssen nun die Massen der Bruchstücke aus dem Massenspektrum ermittelt werden, die aus der Fragmentierung der Kohlenhydratringe resultieren.
Eine detaillierte Beschreibung dieser Methoden findet sich im Kapitel, Entwicklung des Programms *GLYCO-FRAGMENT*.

3.5.2 Glykan-Profiling

Das Glykan-Profiling bezeichnet die Auftrennung und Bestimmung der Glykan-Strukturen, die in einer Zelle oder einem Organ vorkommen. Die Methode kann unter anderem dazu dienen, krankes von normalem Gewebe zu unterscheiden. So zeigen sich im krankhaften Gewebe bei der rheumatoiden Arthritis und dem Rinderwahnsinn deutliche Unterschiede im Glykosylierungsmuster bestimmter Proteine, die darauf hindeuten, dass zelluläre oder genetische Veränderungen die Aktivität der Glykosyltransferasen beeinflussen[51].
Da die Art und Menge der O- und N-Glykane, wie in der Einleitung beschrieben, sehr spezifisch für eine Zelle, sogar für einen Stoffwechselzustand einer Zelle ist, ist es wichtig dieses Profil der Zuckerstrukturen schnell und einfach zu ermitteln. Dazu erfolgt in der Regel zuerst ein chemisches oder enzymatisches Abtrennen der Glykane. Alternativ kann auch vorher das Proteingemisch der Zelle durch Blotting-Techniken oder chromatographische Verfahren[34] aufgetrennt werden. Nach Abtrennen der Protein-Anteile und Aufreinigung der Kohlenhydrate erfolgt eine Trennung der unterschiedlichen Zuckerstrukturen mittels HPLC[35] oder Kapillar-Elektrophorese[52]. Die Detektion der Verbindungen kann einerseits durch die MS[34] geschehen, andererseits auch durch Anbringen eines Chromophors an die Glykane[53] erfolgen, so dass diese Verbindungen auch mit handelsüblichen UV-VIS-Detektoren detektiert werden können. Für die einfache Identifizierung dieser Verbindungen bietet sich eine Suche in einer Datenbank mit Vergleichsspektren an, da eine Standardisierung der HPLC-Bedingungen sehr aufwendig ist, weil schon geringe Unterschiede in den verwendeten Fließmitteln und der verwendeten Säule zu einer Verschiebung der Retentionszeiten im Chromatogramm führen kann. Leider existiert bis jetzt keine Datenbank mit gemessenen Massenspektren. In dieser Dissertation ist daher der Ansatz verfolgt worden, eine Spektral-Datenbank mit berechneten Fragmenten von Strukturen aus der *SWEET-DB* zu erstellen.

3.5.3 Strukturaufklärung von Glykosylierungen

Bei entsprechender Wahl des Ionisierungsverfahrens ist es möglich, dass das zu analysierende Molekül in nicht zu kleine Fragmente zerfällt[46, 54], sondern

hauptsächlich Bruchstücke der glykosidischen Bindungen und zu einem geringeren Anteil cross-ring-Bruchstücke entstehen (siehe auch Abschnitt Nomenklatur der Fragmente). Die Fragmentierung der glykosidischen Bindungen ist schon bei Verwendung von relativ geringen Ionisierungsenergien möglich, da nur eine kovalente Bindung dabei gebrochen werden muss.
Bis zum Ende der neunziger Jahre war die Anregungsmethode der Wahl das Fast-Atom-Bombardment (FAB). Dabei wurde die Ionisierung der zu untersuchenden Probe durch Beschuss mit geladenen Atomen erreicht. Diese Methode hat aber den Nachteil, dass die zu untersuchenden Verbindungen oft in zu kleine Bruchstücke zerfallen[55], so dass Aussagen über die Zusammensetzung der untersuchten Verbindung mit einer großen Unsicherheit behaftet sind. Diese Einschränkungen lassen sich durch Verwendung der Elektrospray-Ionisation (ESI) als Anregungs-methode verringern[56], da diese Methode eine größere Weichheit besitzt und daher Bruchstücke entstehen, die eine höhere Aussagekraft besitzen. Dabei werden mit Hilfe dieser Technik die unterschiedlichsten Fragen der Strukturaufklärung beantwortet. In erster Linie wird mit dieser Methode die Sequenz von Glykanen oder Lipopolysacchariden bestimmt. Sie kann auch für die Bestimmung der Anzahl der Antennen und auch die Art der glykosidischen Verknüpfung eingesetzt werden.

3.6 Methodik

Bei der Analyse von Glykosylierungen von Proteinen wird in aller Regel wie folgt vorgegangen: Nachdem die Proteine isoliert und aufgereinigt worden sind, wird durch Zugabe von Pngase enzymatisch oder durch Hydrazin chemisch die kovalente Bindung gespalten und dadurch das Glykan abgetrennt. Dieses wird nun durch chromatographische oder elektrophoretische Methoden von den Proteinresten abgetrennt und anschließend massenspektrometrisch identifiziert. Zur Erzeugung der Ionen werden als Anregungsmethoden bevorzugt FAB und ESI benutzt. Wobei aber in der letzten Zeit die Anregung durch ESI bevorzugt wird, da so besser interpretierbare Spektren mit höherem Informationsgehalt erzielt werden[57]. Die Nomenklatur der dabei entstehenden Bruchstücke wird in dem nun folgenden Abschnitt beschrieben.

3.6.1 Nomenklatur der entstehenden Fragmente

Die mittels MS erhaltenen Spektren enthalten Fragmente, die entsprechend den allgemein anerkannten Definitionen von Domon und Costello[58] benannt werden. Diese Nomenklatur unterscheidet zwei verschiedene Arten von Fragmenten. Zum einen entstehen Fragmente, die durch das Brechen einer glykosidischen Bindung entstehen. Zum andern bilden sich auch Bruchstücke, die durch das Brechen von zwei Bindungen innerhalb eines Ringes entstehen. Diese werden als Crossring-Bruchstücke benannt. Der Buchstabe A entspricht dabei einem Fragment des nicht-reduzierenden Endes des Kohlenhydrats, das sich durch den Bruch zweier Bindungen eines Zucker-Ringes bildet. Das entsprechende zweite Fragment mit dem reduzierenden Ende wird mit dem Buchstaben X bezeichnet. Zur Charakterisierung der Bindungsbruchstellen wird die Nummer der jeweiligen Bindungen als hochgestellter Index vor den Buchstaben gestellt. Dabei erhält die Bindung zwischen dem Ring-Sauerstoff und dem Kohlenstoffatom C_1 den Wert 0, die Bindung zwischen C_1 und C_2 den Wert 1 und so weiter.
Als tiefgestellter Index wird die Linkage-Information (siehe Absatz 3.6.2) der Residuen in der Zuckerkette geschrieben. Das Fragment $^{1,5}A_3$ entspricht also

dem Bindungsbruch der Bindungen 1 und 5 des dritten Residuums vom nicht-reduzierenden Ende aus gezählt. Die Fragmente, die bei der Spaltung einer glykosidischen Bindung entstehen, werden vom nicht-reduzierenden Ende mit B bezeichnet, wenn die Spaltung vor dem Sauerstoffatom der Bindung erfolgt sonst mit C benannt. Auch hier gibt ein tiefgestellter Index die Position des Residuums in der Zuckerkette wieder. Analog dazu werden vom reduzierenden Ende die Fragmente mit Y, wenn das Sauerstoffatom der Bindung enthalten ist, sonst mit Z bezeichnet. Eine grafische Darstellung dieser Bezeichnungen findet sich in Abbildung 19.

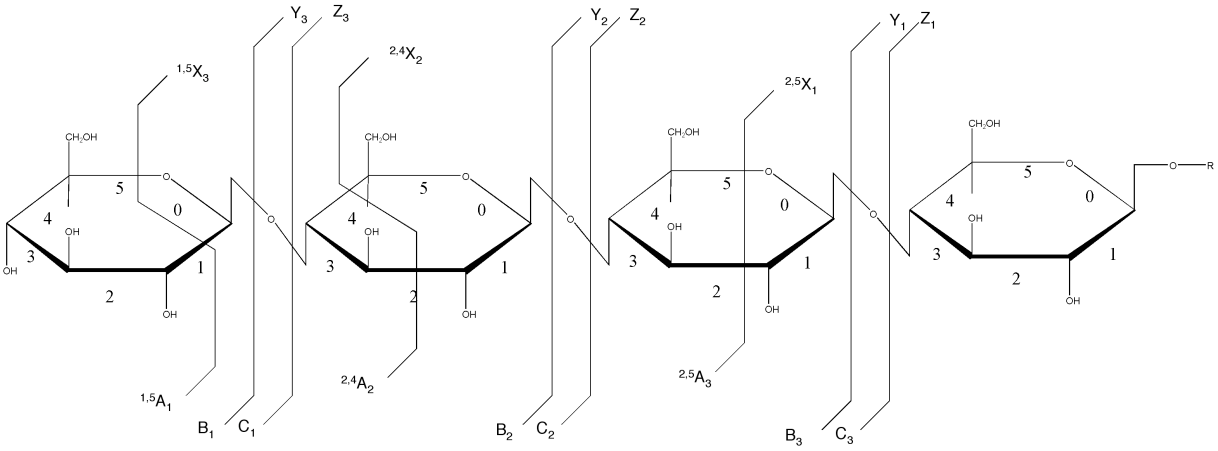


Abbildung 19: Bezeichnung der Fragmente eines Kohlenhydrates nach Domon und Costello

3.6.2 Einführung einer Linkage-Information zur eindeutigen Bezeichnung von Residuen

Die strukturelle Vielfalt der Kohlenhydrate und insbesondere die Möglichkeit, stark verzweigte Topologien zu bilden, machte es nötig eine zusätzliche Bezeichnung einzuführen, die es erlaubt, die einzelnen Residuen eindeutig zuzuordnen. Die von Domon und Costello eingeführte Bezeichnung der Seitenarme mit griechischen Buchstaben wird leider sehr schnell unübersichtlich. Auch ist eine derartige Beschreibung weniger gut für eine Verarbeitung in Computer-Programmen geeignet. Diese Gründe führten zu der Überlegung, als zusätzliche Beschreibung die Linkage-Information bei allen in dieser Arbeit beschriebenen Anwendungen für eine eindeutige Zuordnung der einzelnen Residuen zu verwenden. Entsprechend der IUPAC-Nomenklatur[22, 23] sind die mit dem Programm *SWEET-II* erzeugten Topologien so sortiert, dass das reduzierende Ende einer Kohlenhydratstruktur rechts steht und als erstes Residuum in der internen Liste erscheint. Dieses Residuum wird nun als Startpunkt für die Linkage-Information verwendet. Ausgehend davon werden immer die Kohlenstoff-Atome benannt, die damit verbunden sind. In Abbildung 20 ist beispielhaft das Residuum □-D-GalpNAc (roter Pfad) ausgewählt, und besitzt die Linkage-Information 3,3,4,3.

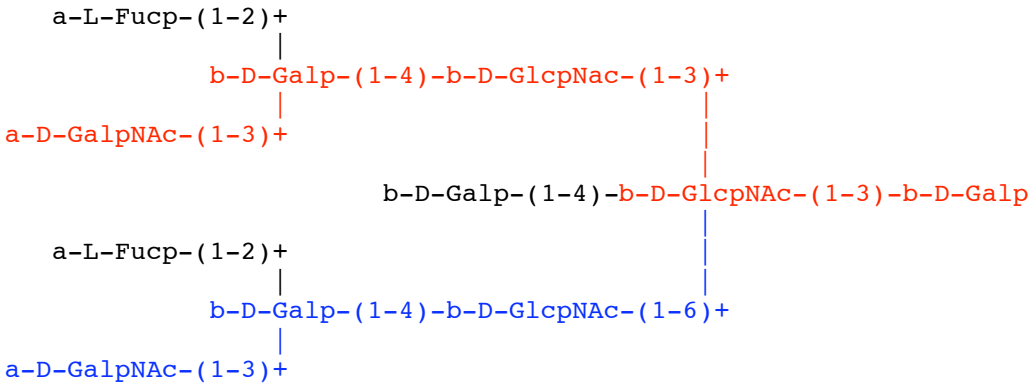


Abbildung 20: □-D-GalpNac mit der Linkage-Information 3,3,4,3

Mit der Angabe der Linkage-Information ist ein Residuum in einer Verbindung eindeutig bezeichnet und kann nicht mit dem zweiten endständigen □-D-GalpNAc-Residuum (blauer Pfad) verwechselt werden, das die Linkage-Information 3,6,4,3 besitzt. So kann selbst bei Verbindungen, wie z.B. High-Mannose Glykanen, sehr einfach und schnell auf einzelne Residuen verwiesen werden, und diese können auch sehr schnell identifiziert werden.
Die Bruchstücke, versehen mit der Linkage-Information als Index, lassen sich nun gut als Bezeichnung in den Spektren verwenden. Dieses wird an dem folgenden Beispiel verdeutlicht:

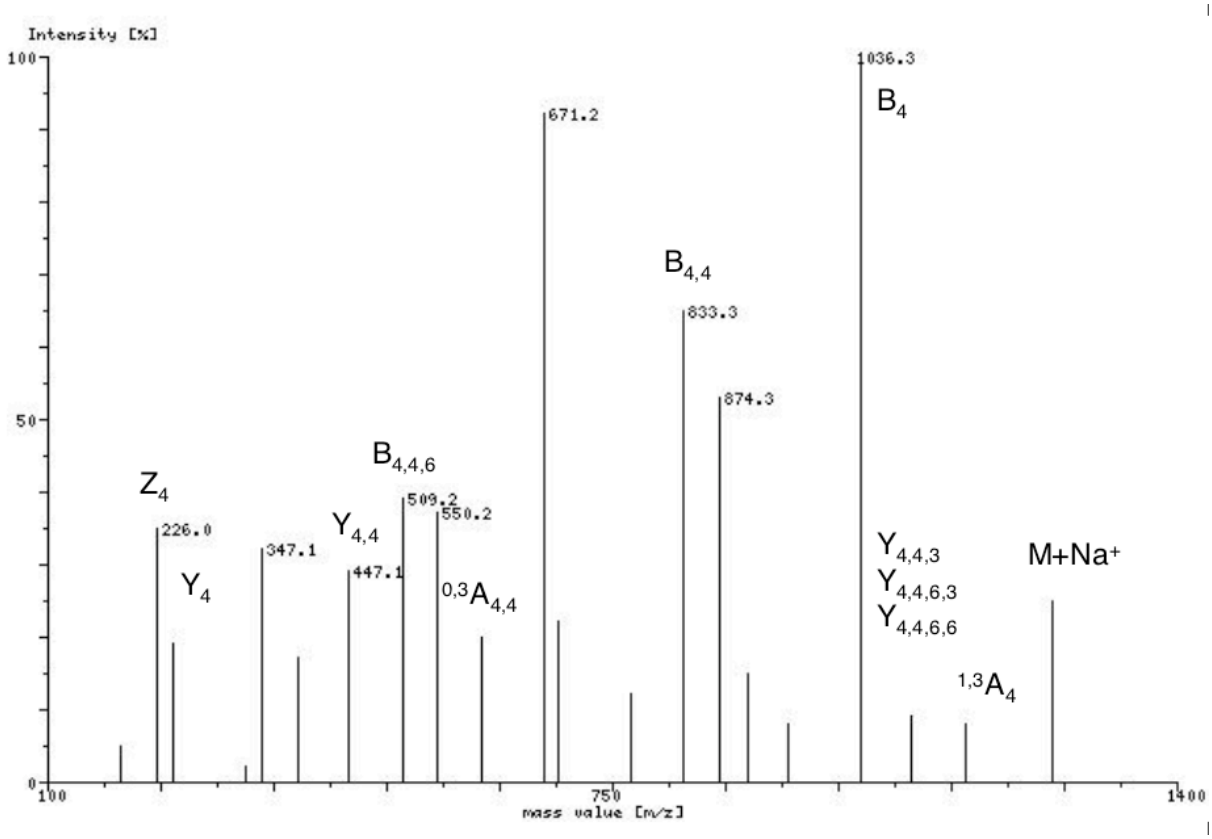


Abbildung 21: Bezeichnung der Fragmente mittels der Linkage-Information in einem Massenspektrum eines Kohlenhydrats

3.7 NMR Untersuchungen von Kohlenhydraten

Mittels der Kernresonanzspektrometrie (NMR) kann zusätzlich zur MS auch die Art und Konfiguration der monomeren Einheiten, die Art der glykosidischen Verknüpfung

und damit auch die exakte Sequenz von Kohlenhydraten ermittelt werden[25, 26]. Zusätzlich können NMR Experimente detaillierte Daten zur Abschätzung der Konformationen der Glykane liefern. Allerdings liegt die benötigte Menge an reiner Substanz um Größenordnungen, für gute Spektren im Milligramm-Bereich, höher als bei den MS-Verfahren, so dass NMR-Techniken weniger gut geeignet sind für Hochdurchsatzverfahren. NMR-Messungen bieten sich in der Regel zur Identifizierung eines Kohlenhydrats durch Vergleich mit einem Referenzspektrum an[59-61]. Diese Referenzdaten müssen dem Benutzer aber auch zum Vergleich zur Verfügung gestellt werden. Dieses macht es nötig, die Spektren in Datenbanken zum Vergleich bereitzuhalten.

3.8 Gewinnung, Bereitstellung und Verbreitung von wissenschaftlichen Daten im Zeitalter des Internets

Die Bioinformatik ist zu einem integralen Bestandteil der bio-medizinischen Forschung geworden. Eine ständig steigende Anzahl von Forschern auf der ganzen Welt nutzt intensiv Sequenzdatenbanken und bioinformatorische Werkzeuge zur Auswertung der eigenen gemessenen Rohdaten. Die dabei entstehenden Datenmengen werden immer umfangreicher und müssen durch (bio)informatorische Ansätze sinnvoll verarbeitet werden. So werden immer mehr Genome entschlüsselt, die Struktur von immer mehr Proteinen wird aufgeklärt und die Automatisierung experimenteller Methoden zur Untersuchung in der Biotechnologie wird ständig vorangetrieben. Dieses erklärt auch die steigende Anzahl der Einträge in der *BROOKHAVEN-PDB*[62], einer zentralen Datenbank in der räumliche Strukturen von Proteinen abgelegt werden. Diese neu entstandenen Daten und Informationen müssen den Forschenden und einem interessierten Publikum aber auch wieder zugänglich gemacht werden. Dazu bieten sich zwei verschiedene Verfahren an. Zum einen kann dies durch die Verbreitung von Daten und Programmen als so genannte Standalone-Anwendung (Einzelplatzanwendung) erfolgen. Die Alternative dazu sind webbasierte Lösungen, bei denen die Programme und Daten zentral auf einem Server zur Verfügung gestellt werden, die mittels Web-Browser über das Internet abgerufen werden können.

	Einzelplatzanwendung	Webbasierte Lösung
Installation	Kann sehr aufwendig sein	Nicht nötig, da alle Komponenten in der Regel schon installiert
Betriebssystemabhängigkeit	Ja	Nein
Geschwindigkeit	Gute Ausführungsgeschwindigkeit	Durch Browser Geschwindigkeitseinbussen
Datensicherheit	Bei Trennung vom Internet ja	Nein
Informationsdarstellung	Keine Grenzen	Auf die Möglichkeiten des Browsers und der Plugins beschränkt
Interaktion mit dem Benutzer	Gut	Bei langsamen Antwortzeiten des Servers eher schleppend

Tabelle 4: Vor- und Nachteile von webbasierten Lösungen und Einzelplatzanwendungen

Die beste Lösung verspricht sicher eine Verschmelzung der beiden Lösungen. So kann je nach Anforderung des Algorithmus auch eine Einzelplatzanwendung die ideale Lösung sein, die über das Internet mit einem Server kommuniziert.

3.8.1 Datenbanken

Viele Datenbanken in der Bioinformatik verfolgen mittlerweile einen webbasierten Ansatz. Die umfangreichste Quelle von Daten zu Proteinen die *SWISSPROT*[63] und *TREMBL*-Datenbank[64] des Schweizer Instituts für Bioinformatik sind webbasierte Datenbanken, die es ermöglichen, eine Vielzahl von gut annotierten Daten und Informationen zu Proteinen zu erhalten. Solange die Datenausgabe mittels Standard-Webbrowsern erfolgen kann und keine speziellen grafischen Darstellungsoptionen verlangt werden, überwiegen die Vorteile des webbasierten Ansatzes. Viele Datenbestände weiterer Datenbanken werden ebenfalls in dieser Form angeboten. Einen guten Überblick erhält man durch das aktuelle Heft des Journals ‚Nucleic Acids Research'[65]. Sehr bekannt ist auch die *BROOKHAVEN PROTEIN DATABANK*, die Röntgenstrukturen für Proteine enthält. Eine der wichtigsten Datenbanken im Bereich der Glykobiologie war die *CARBBANK*, die Strukturdaten und taxonomische Daten enthielt. Eine übergeordnete Stellung nimmt die *PUBMED* ein. Diese ist eine zentrale Datenbank in der sich die Publikationsdaten der wichtigsten wissenschaftlichen Journale befinden. Eine detaillierte Beschreibung der einzelnen Datenbanken befindet sich im Kapitel ‚Untersuchungen zur automatischen Pflege und Annotierung einer Datenbank'. In diesem Kapitel werden auch die Vor- und Nachteile der einzelnen Methoden zur Aktualisierung und Pflege der Daten besprochen.

3.8.1.1 Pflege und Annotierung einer Datenbank

Ein sehr wichtiger Prozess bei der Verwaltung einer Datenbank ist der Prozess der Annotierung und der Neueingabe der Daten. Dieses ist jedoch ein sehr empfindlicher Prozess für eine Datenbank. So wird sie für den Benutzer unbrauchbar, sobald sich zu viele fehlerhafte oder nicht in das Themengebiet gehörende Daten darin befinden. Dieses erhöht aber den Zeitbedarf und die Kosten der Datenbank nicht unerheblich.

Dabei werden verschiedene Strategien verfolgt, die alle ihre Vor- und Nachteile besitzen. Eine Übersicht dazu ist in der folgenden Tabelle enthalten.

Strategie	Vorteile	Nachteile
Zentrale Datenbank mit zentraler Eingabe	Die Daten sind in sich konsistent und von einer hohen Qualität.	Durch manuelle Extraktion und Eingabe sehr teuer und auch sehr langsam
Zentrale Datenbank mit automatischer Eingabe	Preiswerte und auch sehr schnelle Eingabe	Keine Überprüfung der gefundene Daten Keine Rohdaten
Zentrale Datenbank mit Eingabe der Daten vor Ort	Durch die vorhandenen Rohdaten besitzen die Daten eine hohe Qualität Gute Datenkonsistenz	Geringe Vernetzung und Kooperation der Forschungsgruppen
Lokale Datenbank mit Eingabe der Daten vor Ort	Preiswerte Verwaltung der Daten Hohe Akzeptanz	Keine Qualitätskontrolle

Tabelle 5: Strategien zur Pflege und Verwaltung von Datenbanken

In der Vergangenheit sind diese Dinge weitgehend manuell und auch zentral vorgenommen worden, erst mit der Entwicklung des Internets kann dieses auch dezentral erfolgen. In den folgenden Abschnitten soll nun untersucht werden, wie sich die Annotation der Daten durch den Einsatz von Programmen deutlich effektiver gestalten lässt. Im Moment wird in Amerika durch das Consortium for Functional Glycomics[1] versucht, eine Datenbank für Informationen zu Kohlenhydraten, wie taxonomisches Vorkommen, Struktur, chemische und physikalische Eigenschaften und deren physiologischen Funktionen zu schaffen. Es gibt jedoch auch viele kleinere bis kleinste Datenbankprojekte. Eine Auswahl davon wird in den folgenden Abschnitten dargestellt. Es sollte allerdings darauf geachtet werden, dass diese Datenbanken untereinander quervernetzt sind, da der Benutzer so schnell von einer Datenbank zur anderen gelangen kann.
Eine weitere Möglichkeit, den Datenbestand einer Datenbank zu erhöhen, besteht in der Möglichkeit aus den vorhandenen Daten neue abzuleiten. So kann zum Beispiel aus einer Proteinsequenz ohne größeres Problem die chemische Summenformel oder die Masse abgeleitet werden und als zusätzliche Information bereitgestellt werden.

3.8.1.2 Data-Warehousing

Unter dem Begriff Data-Warehousing werden Strategien zusammengefasst, die sich mit der Bereitstellung von Informationen aus verschiedenen Internet-Quellen befassen. Dazu werden in einer übergeordneten Datenbank verschiedene Suchmöglichkeiten für die gespeicherten Daten angeboten. Für die detaillierteren Ergebnisse wird aber die Bereitstellung der Daten von einem anderem Server

[1] http://web.mit.edu/glycomics/consortium/

erledigt. Man hat so die Möglichkeit, die Datenbanken an verschiedenen Stellen zu pflegen, sie aber trotzdem unter einer gemeinsamen Oberfläche anzubieten.

3.8.1.3 Strategien zur automatischen Informationsextraktion (Data mining)

Die Kosten und die Zeit für die Eingabe der Daten haben in der näheren Vergangenheit dazu geführt, dass Anstrengungen unternommen worden sind, automatisch Informationen aus Texten zu extrahieren. Dies sind in der Regel allerdings Informationen, die in Textform vorliegen, wie z. B. die Namen von Arzneistoffen und Geninformationen aus einer Textstelle[66]. Es gibt aber auch Versuche eine bestimmte Anzahl von Texten zu analysieren, und die darin enthaltenen Informationen zu organisieren[67].

3.8.2 Web-Applikationen

Bis jetzt wurde nur auf Datenbanken hingewiesen mit denen es möglich ist, eine Datenrecherche über das Internet vorzunehmen. Eine andere Art von Webservices sind Webapplikationen, die es gestatten, Daten über das Internet auf einem Server zu bearbeiten. Sehr gute Beispiele für dieses Vorgehen sind die Programme *SWEET-II*[68, 69] und *PDB2MGIF* von Andreas Bohne-Lang aus unserer Abteilung. Mit dem Programm *SWEET-II* lässt sich z.B. aus einer Kohlenhydratsequenz eine dreidimensionale Struktur berechnen. Dazu wird die Sequenz in einer IUPAC konformen Nomenklatur in ein Webformular eingetragen. Das Programm auf dem Webserver berechnet die Koordinaten für die dreidimensionale Struktur. Diese Ergebnisse werden dann lokal beim Benutzer mit einem speziellen Visualisierungs-programm dargestellt.

3.9 Ausgangssituation

Angesichts der zu erwartenden Datenmengen, die im Rahmen sich abzeichnender Glykomik-Projekte anfallen werden, ist es dringend notwendig, dass auch entsprechende bioinformatorische Werkzeuge und Datenbanken für die Glykobiologie entwickelt werden. Aufgrund der Unterschiede in ihren strukturellen Merkmalen lassen sich die zur Beschreibung von Ähnlichkeiten und Eigenschaften von DNA- und Proteinsequenzen entwickelten Algorithmen nicht einfach übertragen. Dabei muss man zwei grundsätzliche Dinge unterscheiden. Zum einen sind es Algorithmen, die es dem Forscher erleichtern, seine Rohdaten besser auszuwerten. Zum anderen muss der Forscher den Überblick über die aktuelle Forschung behalten. Dazu ist es nötig, Datenbanken bereit zu halten, die zum einen Rohdaten als Vergleichsdaten enthalten, aber auch die aktuelle Literatur zu den unterschiedlichen Strukturen anbieten. Es müssen also Strategien entwickelt werden, die es gestatten, die Messergebnisse, die mit unterschiedlichen experimentellen Ansätzen und Methoden von verschiedenen Forschern über eine Datenstruktur verlinkt sind, zusammenzufassen und mittels eines Webinterfaces diese zur Verfügung zu stellen. Es kann versucht werden, aus den vorhandenen Daten neue abzuleiten. Außerdem sollte auch versucht werden, dass die aktuelle wissenschaftliche Literatur gescreent wird. Dabei sollten Daten, die in den Bereich der Glykobiologie gehören, extrahiert werden. Dies ist in den letzten Jahren verstärkt geschehen. Eine Auswahl der bioinformatorischen Lösungen findet sich in den folgenden Abschnitten.

3.10 Einsatz von webbasierten Softwarewerkzeugen zur Sequenz- und Strukturaufklärung

Zur Auswertung von Massenspektren von Kohlenhydraten fehlt es an entsprechenden Softwaretools. Es sind viele Forscher im Bereich der MS damit beschäftigt, die Bruchstücke nach Domon und Costello mittels eines Taschenrechners zu bestimmen, oder häufig auftretende Peaks durch Expertenwissen wieder zu erkennen. Erste Ansätze dieses automatisch zu erledigen, wurden mit den Tools *GLYCOMOD* und *GLYCANMASS* zur Verfügung gestellt.
Am weitesten fortgeschritten ist die Sequenzbestimmung von Proteinen durch das Programm *MASCOT*, da sich hier für den Computer ideale Bedingungen ergeben. Zum einen sind hier die zu ermittelnden Sequenzen linear angeordnet und zum andern gibt es eine große Anzahl von existierenden und annotierten Datenbanken mit Sequenzinformationen, die zur automatischen Verarbeitung zur Verfügung stehen. Die Strukturvielfalt der Kohlenhydrate gestaltet diese maschinelle Auswertung bedeutend schwieriger. Es kann z. B. nicht einfach auf einen Einbuchstabencode zurückgegriffen werden. Bisher sind nur sehr unbefriedigende Versuche unternommen worden, die Strukturaufklärung von Kohlenhydraten und die Auswertung von Massenspektren mit Hilfe von Programmen und Webapplikationen zu vereinfachen.

3.11 Aufgabenstellung

Bei der Auswertung des Genoms und Proteoms haben Automatisierung und die Entwicklung von intelligenten Softwarewerkzeugen einen wesentlichen Beitrag geleistet. Sie ermöglichen es, die anfallende Datenmenge sinnvoll zu verwalten und auszuwerten. Es wird geschätzt, dass die Entschlüsselung des humanen Genoms durch die bioinformatorische Entwicklung um ungefähr zwei Jahre verkürzt worden ist[70].
Die Entwicklung von Datenbanken und Bioinformatik-Werkzeugen für die Glykobiologie befindet sich noch in den Kinderschuhen[71]. Es ist jedoch offensichtlich, dass die jetzt entstehenden Glykomik-Projekte eine starke Nachfrage erzeugen werden. Bei dem größten bisher laufenden Projekt, dem amerikanischen *Consortium for Functional Glycomics*, ist die Entwicklung einer Bioinformatik-Infrastruktur ein ganz zentraler Punkt. Momentan wird intensiv daran gearbeitet, drei neue Datenbanken für Glykosyltransferasen, Lektine und Kohlenhydrate aufzubauen. Die Übertragung der Algorithmen aus dem Bereich der Genomik und Proteomik auf den Bereich der Glykomik ist aber nicht ohne weiteres möglich, da die strukturelle Vielfalt der Kohlenhydrate aufwendigere Ansätze nötig macht. Hauptziel dieser Arbeit ist es zu untersuchen, inwieweit es mit Hilfe von bioinformatorischen Ansätzen möglich ist, die automatische Analyse von Messergebnissen in der Glykobiologie zu erleichtern. Das Hauptaugenmerk wird dabei auf die Auswertung von Massenspektren gelegt, da sich klar abzeichnet, dass sich die MS als Standardmethode für die schnelle Analyse von Glykomen etablieren wird. Arbeitserleichterung und erhöhte Effizienz für die praktische tägliche Arbeit der Forschenden sind wesentliche Ziele des Designs der neu zu entwickelnden Anwendungen und Datenbanken.
Ein zweiter wesentlicher Aspekt dieser Arbeit ist, zu untersuchen in wie weit es unter der in Tabelle 5 genannten Strategien möglich ist, das Einspielen von neuen Einträgen in die Datenbank *SWEET-DB*,[59] zu automatisieren. Als externe Quellen stehen einerseits die digital allgemein verfügbare Referenz-Datenbank *PUBMED* zur

Verfügung und anderseits sind es die jeweiligen Originalpublikationen, die heute auch meist in digitaler Form zugänglich sind. In Zukunft soll es möglich sein, Literatur, Rohdaten und aus diesen errechnete Daten durch effiziente Suchalgorithmen und geeignete Methoden in der Datenbank zu speichern. Dieser Prozess soll nach Möglichkeit vollautomatisch, aber auch durch direkte Eingabe von Daten aus Publikationen und Programmen, die von den Benutzern lokal auf ihren Festplatten gespeichert sind, erfolgen können.
Des Weiteren ist es aber auch nötig, dem Benutzer Wissen zur Verfügung zu stellen, dass durch Data-Warehousing aus anderen frei zugänglichen Quellen gewonnen worden ist. So kann durch Klassifizierung von Publikationen eine Vorauswahl erfolgen, die bei der Recherche zu einer nicht unerheblichen Beschleunigung des Informationsgewinns führt. Es sollte also nicht mehr nötig sein, auch thematisch nicht passende Informationen zu sichten. Sollte es zum Beispiel erforderlich sein, dass man Informationen für ein bestimmtes N-Glykan sucht, reicht es, dieses in der *SWEET-DB* zu suchen und sich die dazu gehörige Literaturliste anzusehen, da automatisch alle neuen Publikationen zu dem N-Glykan eingetragen werden. In dieser Arbeit sollen dazu bioinformatorische Ansätze untersucht werden mit denen Daten in die *SWEET-DB* geschrieben werden können. Alle Applikationen und Datenbanken sollen das Internet als Basis für den Datenaustausch und die Bereitstellung der Informationsangebote besitzen.

4 Entwicklung von Softwarewerkzeugen zur Analyse von Massenspektren

4.1 Allgemeines

Die MS hat sich seit ca. 1985 zu einer wichtigen Analysemethode für die Untersuchung von Proteinen und Kohlenhydraten entwickelt. Es begann damit, dass mittels der Entwicklung der MALDI-Technik das Molekulargewicht ganzer Proteine bestimmt werden konnte. Es stellte sich aber auch sehr schnell heraus, dass mit dieser Technik auch kleinere Moleküle sehr schonend untersucht werden konnten. So wurden bereits im Jahre 1987 erste Kohlenhydrate mit Hilfe der MS untersucht[72]. Zur Ionisation wurde am Anfang sehr häufig das Fast-Atom-Bombardment (FAB) verwendet[73]. Damit war in vielen Fällen eine ausreichende Analyse der zu untersuchenden Substanzen möglich. Es musste jedoch, um aussagekräftige Spektren zu erhalten, sehr häufig eine Permethylierung erfolgen[74-76]. Mit Entwicklung der Elektrospray-Ionisation (ESI) wurde dann eine noch schonendere Ionisationsmethode entwickelt, die es gestattete auch labilere Moleküle wie die Substanzklasse der Kohlenhydrate zu untersuchen, ohne das eine Permethylierung durchgeführt werden musste[39, 77].
Die MS ist eine vielfältig einsetzbare Technik mit der auf standardisiertem Wege sehr schnell gut reproduzierbar Moleküle identifiziert werden können. In den folgenden Abschnitten soll nun untersucht, wie durch den Einsatz von (bio)informatorischen Ansätzen die Auswertung der Spektren, insbesondere von Kohlenhydraten, vereinfacht und automatisiert werden kann.

4.2 Signalarten

Der Verteilung der Isotopen einer Atomart kann einen entscheidenden Einfluss auf die Signale des gemessenen Spektrums haben, da es für fast jedes in der Natur vorkommende Element unterschiedliche Isotope gibt. Diese führen bei der Analyse eines Moleküls zu einer Massenverteilung der einzelnen Fragment-Ionen, die genau dem Verteilungsmuster dieser Isotopen entsprechen.
Das Isotopenmuster eines Moleküls lässt sich aus der Binominalverteilung der im Molekül enthaltenen Atomarten ermitteln.

$$Isotopenverteilung = \left(I_{X_c} + I_{Y_c} + I_{Z_c} + ...\right)^{n_c}\left(I_{X_H} + I_{Y_H} + I_{Z_H} + ...\right)^{n_c}\left(I_{X_o} + I_{Y_o} + I_{Z_o} + ...\right)^{n_o}...$$

wobei C, H, O den Elementen entspricht
X, Y, Z den Isotopenmassen entspricht
I_{XC}, I_{YC}, I_{ZC} der Isotopenhäufigkeit entspricht
n_c, n_h, n_o der Anzahl der jeweiligen Atome entspricht

Ausgehend von dieser Formel kann die Masse eines Moleküls je nach der Isotopenverteilung berechnet werden. Bei Peptiden und Kohlenhydraten, die eine Masse kleiner als 5KDa haben, wird in der Regel die monoisotopische Masse angegeben, da Wasserstoff, Kohlenstoff und Sauerstoff zu mehr als 99 Prozent aus nur einem Isotop bestehen. Diese Massen können sehr gut von den heutigen Massenspektrometern mit einem Fehler von deutlich unter 0,02 Da gemessen werden. Bei Massen, die größer als 5KDa sind,wird aber in der Regel die

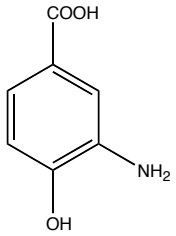
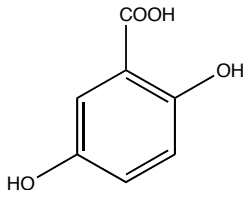
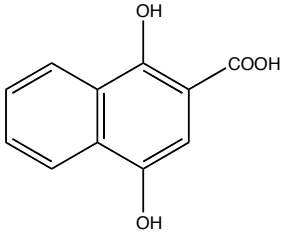
durchschnittliche Masse angegeben, da in diesen Massenbereichen die einzelnen Isotopenpeaks nicht mehr aufgelöst werden können.

4.3 Fast-Atom-Bombardment (FAB)

Bei dieser Ionisierungstechnik wird mittels stark beschleunigter Primärteilchen, wie den Edelgasen Argon oder Xenon, aus der zu untersuchenden Lösung ein konstanter Ionenstrahl gewonnen[78]. Die Substanz muss dazu in einer oberflächenaktiven, flüssigen und schwerflüchtigen Matrix (z. B. Glycerin) vorliegen. Durch Zusatz von Alkalisalzen entstehen bevorzugt metallkationische Ionen (Na^+, K^+). Der Strom von Sekundärteilchen wird anschließend durch eine angelegte Spannung beschleunigt und kann im Massenspektrometer untersucht werden. Diese Technik gehört schon zu den weicheren Ionisierungstechniken. Je weicher eine Ionisierungstechnik ist, desto weniger wird das zu untersuchende Molekül dabei zerstört. Für die Analyse von Proteinen und Kohlenhydraten sind schonende Ionisierungstechniken sehr wichtig, da für eine effektive Sequenz- bzw. Strukturaufklärung möglichst nur solche Fragmente entstehen sollen, die den bekannten Bindungsbrüchen entsprechen.

4.4 MALDI

Grundlage bei der Matrixunterstützte Laserdesorption/Ionisation (MALDI) ist der konstante Laserbeschuss einer festen, kristallinen und UV-Strahlung absorbierenden Matrix, in der das zu untersuchende Protein oder Kohlenhydrat eingebettet ist. Diese Matrix muss aus einer niedermolekularen, leicht zu kristallisierenden Substanz bestehen, die durch den verwendeten Laser energetisch angeregt werden kann. Für die Analyse von Kohlenhydraten typische Matrixsubstanzen sind in Tabelle 6 gezeigt.

Derivativ	Abkürzung	Struktur
3-Amino-4-Hydroxybenzoesäure	-	
2,5-Dihydroxybenzoesäure	2,5 DHB	
1,4-Dihydroxy-Naphthalsäure		

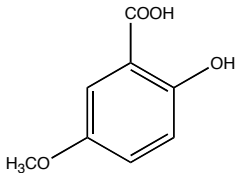
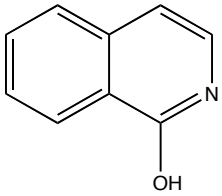
2-Hydroxy-5-Methoxy-Benzoesäure	-	
1-Hydroxychinolin	1-HIQ	

Tabelle 6: Liste der gebräuchlichsten Matrixsubstanzen

Eine wesentliche Aufgabe dieser Matrix ist es, dass ein Teil der Energie des Lasers absorbiert wird. Diese gibt dosiert die absorbierte Energie an die eingebettete zu untersuchende Substanz weiter und bewirkt, dass sie in nicht zu kleine Bruchstücke zerfällt. Die Methode eignet sich auch zur Massenbestimmung sehr großer Moleküle, so dass sie als erstes zur Bestimmung des Molekulargewichtes von Proteinen eingesetzt worden ist.

4.5 Elektrospray-Ionisation

Bei der Elektrospray-Ionisation (ESI) wird eine Lösung des Analyten zusammen mit einer Elektrolytlösung mittels sehr feinen Kapillaren in ein starkes elektrisches Feld versprüht. Die Spannung dieses Feldes ist dabei zwischen Kapillarspitze und einer Gegenelektrode angelegt und ist verantwortlich für eine effektive Ionisierung des Analyten. Es können je nach Polarität der angelegten Spannung positive und auch negative Ionen gemessen werden. Als Elektrolyten werden für Messungen im positiven Modus Alkali-Ionen verwendet. Dabei erhält man mit Verwendung von Lithium die intensivsten Ionen[57].
Im Gegensatz zur FAB-Ionisierung ist diese Methode sehr schonend und führt bei der Analyse von Zuckern dazu, dass hauptsächlich Ionen, die aus einem Bruch der glykosidischen Bindungen und eine bestimmte Anzahl von sehr charakteristischen Ringfragmenten[79] resultieren, die Aussagen über die Sequenz der untersuchten Struktur zulassen. Ein weiterer Vorteil dieser Methode ist eine bedeutend bessere Sensitivität für die Analyse von Glykoproteinen[80], Gangliosiden[81] und N-Glykanen[82].

4.6 Einschränkungen der MS

Obwohl die MS eine sehr effiziente Methode zur Bestimmung sowohl der Komposition und auch teilweise der Topologie eines Zuckers ist, so besitzt die Methode auch ihre Grenzen. So kann man keine Aussagen über die Stereochemie der untersuchten Zucker-Residuen machen. Bei einem Peak von der Masse 162Da kann nicht unterschieden werden, ob es sich um eine □-D-Glucose, eine □-L-Mannose oder irgendeine andere Hexose handelt. Daher kann mit Hilfe der MS allein nur die Komposition einer Verbindung bestimmt werden. Das dieses schon sehr hilfreich sein kann, zeigt ein Blick in die *SWEET-DB*: Sucht man nach einer Verbindung, die neun Hexosen, eine Pentose und zwei HexAmino-Residuen enthält, so erhält man nur einen Treffer, was für eine weitere, detaillierte Untersuchung der Verbindung sehr hilfreich sein kann.

4.7 Computerbasierte Massenspektreninterpretation

Die computerbasierte Spektreninterpretation von MS/MS-Spektren von Peptiden, die durch den Verdau von Proteinen mit dem Enzym Trypsin entstanden sind, stellt gegenwärtig die am weitesten fortgeschrittene Technologie zur automatischen Identifizierung von Proteinen dar. In vielen Fällen ist eine vollautomatische Identifizierung von Proteinen möglich. Als quasi Standard kann das Programm *MASCOT* der Firma Matrixscience angesehen werden. Die Bindungen des Rückgrades von Peptiden brechen, wie in der Einleitung beschrieben, bevorzugt an bestimmten Bindungen und zeigen somit charakteristische Fragmentierungsmuster, die sich effizient in Algorithmen umsetzen lassen. Eine große Zahl an Proteinsequenzen sind bekannt und ihre Sequenzen in Datenbanken abgelegt, die allgemein verfügbar sind.

4.7.1 MASCOT

Mit dem Softwarepaket *MASCOT* [83][1] von der Firma Matrixscience ist es möglich die Auswertung eines MS- oder eines MS/MS-Spektrums eines Peptids stark zu beschleunigen. Das Programm ist in der Lage auf der Grundlage der angeschlossenen Protein-Datenbank, wie *SWISSPROT*[84] oder der *PIR* Datenbank[85], alle Proteinsequenzen *in silico* zu verdauen. Aus den erhaltenen Peptiden werden dann deren theoretische Massenspektren berechnet. Diese werden dann nacheinander mit dem eingegebenen Spektrum verglichen, und es wird ein Wahrscheinlichkeitsfaktor berechnet, der angibt, wie gut das gemessene und das theoretische Spektrum zueinander passen. Mit diesem Verfahren ist eine sehr effektive und schnelle Methode gegeben, aus dem gemessenen MS/MS-Spektrum eines Peptids die Sequenz zuzuordnen.

4.7.2 GLYCOMOD

Ein erster Versuch die Analyse von Glykanen zu automatisieren ist die Anwendung *GLYCOMOD*[86][2]. Dabei ist es möglich, an Hand einer Liste von Massenwerten die mögliche Sequenz eines Oligosaccharides zu erschließen. Hierbei kann ein beliebiges Esi-Ion ausgewählt werden. Je nach Aufnahmeart des Spektrums kann ausgewählt werden, ob die monoisotopische Masse oder die durchschnittliche Masse berücksichtigt werden soll. Die Proteinsequenz, an die das Glykan gebunden ist, kann ebenfalls eingeben werden. Als Ergebnis erhält man alle Kompositionen, die einem eingegebenen Molpeak entsprechen.

4.7.3 GLYCANMASS

Das webbasierte Tool *GLYCANMASS*[3] berechnet die Masse der Verbindung auf der Grundlage der eingegebenen Komposition und der an dem Glykan befindlichen Modifikationen. Als Berechnungsoptionen kann gewählt werden, ob die durchschnittliche oder monoisotopische Masse berechnet werden soll. Die bei der Analyse von Glykanen häufig angewendete Permethylierung und Peracetylierung kann angegeben werden und wird automatisch für die Verbindung berücksichtigt. Die Massenwerte müssen allerdings für jede Verbindung einzeln berechnet werden.

---

[1] http://www.matrixscience.com

[2] http://us.expasy.org/tools/glycomod/

[3] http://us.expasy.org/tools/glycomod/glycanmass.html

4.7.4 STAT

Bei dem Programm STAT[87][1] handelt es sich um ein Webtool mit dem es möglich ist, die Sequenz eines Kohlenhydrates aus einem MS^n–Spektrum zu ermitteln. In einem ersten Schritt ist es dazu nötig, das Precursor-Ion, das Esi-Ion, die Fehlergrenzen und die Residuen einzugeben, die in der Probe enthalten sind. Ist dieses geschehen, kann in einem zweiten Schritt die Liste der Massen des zu analysierenden sequentiellen Spektrums und auch die mögliche Komposition der Probe eingegeben werden. Im dritten und letzten Schritt wird nun den Massen des eingegebenen Spektrums eine theoretisch mögliche Komposition zugeordnet. Danach erfolgt die Ausgabe der Sequenzen, die diesen Kriterien entsprechen. Dabei wird das beste Ergebnis im Browser zuerst ausgegeben.

4.8 Aufgabenstellung

Ausgehend von dem gegenwärtigen Stand der experimentellen und methodischen Entwicklung der MS-basierten Methoden zur Analyse von Glykanen und den in der Einleitung dargestellten wissenschaftlichen Fragestellungen im Bereich der Glykomik soll nun untersucht werden, welche Möglichkeiten es gibt, die einzelnen Arbeitsschritte durch passende Software-Werkzeuge stark zu beschleunigen oder sogar vollständig zu automatisieren. Der Schwerpunkt bei der Entwicklung von Software-Werkzeugen wird dabei zum einen auf die Positionsbestimmung einer post- oder co-translationalen Modifikation gelegt, zum anderen auf die anschließende Strukturaufklärung der verknüpften Glykane.

4.9 Entwicklung des Programms *FINDYSERIES*

Wie in den vorhergehenden Absätzen beschrieben, ist es wichtig, die Sequenz eines Peptids und die Position von möglichen co- und posttranslationalen Modifizierungen zu finden. Es wird zunächst untersucht wie verfügbare bioinformatorische Ansätze, die eine automatische Interpretation von MS/MS-Spektren von Peptiden erlauben, genutzt werden können, um den experimentellen Massen die *in silico* berechneten Fragmente zuzuordnen. Im Idealfall, unter optimal gewählten Messbedingungen, sollten Serien von a-, b-, x- und y-Ionen zu sehen sein. Je höher die Übereinstimmung zwischen berechneten und experimentell gefundenen Massen ist, umso wahrscheinlicher ist, dass die postulierte Sequenz richtig ist. Ist das untersuchte Fragment modifiziert, so sollten auch Serien zu sehen sein, die um die in Tabelle 3 angeführten Inkremente der jeweiligen Modifikationen erhöht sind. Im Idealfall entspricht das Massenspektrum eines Peptids der Sequenz DICSVTCGGGVQK, das an dem Threonin an der sechsten Position fukosyliert ist[46], einem Spektrum, das die folgenden Peaks enthält. In dem Spektrum ist ein gefundenes Fragment-Ion mit einem Stern (*) gekennzeichnet. Sollte es sich um ein Fragment-Ion handeln, das um den Massenwert der Modifikation erhöht ist, so ist das Vorhandensein des Fragment-Ions mit einem Pluszeichen (+) gekennzeichnet.

[1] http://www.cchem.berkeley.edu/mswww/

#	a-Ion	b-Ion	Sequence	y-Ion	#
1	*	*	D	+	13
2	*	*	I	+	12
3	*	*	C	+	11
4	*	*	S	+	10
5	*	*	V	+	9
6	+	+	T	+	8
7	+	+	C	*	7
8	+	+	G	*	6
9	+	+	G	*	5
10	+	+	G	*	4
11	+	+	V	*	3
12	+	+	Q	*	2
13	+	+	K	*	1

* means peak is present, + peak is modified with Fucose

Tabelle 7: Liste der theoretischen Fragmente der fucosylierten Peptidsequenz DICSVTCGGGVQK

Es ist hier sehr schön zu sehen, dass die a- und b-Ionen ab dem Punkt der Modifikation auch um den Massenwert der Fucose erhöht vorkommen. Dasselbe ist auch bei den y-Ionen zu beobachten. Dies ist ein klarer Beleg dafür, dass die Aminosäure Threonin an der Position sechs fucosyliert ist.

4.9.1 Anforderungen an das Programm *FINDYSERIES*

Das zu entwickelnde Programm sollte möglichst intuitiv zu bedienen sein, so dass Glykowissenschaftler ohne intensives Lesen von Handbüchern alle Funktionen des Programms aufrufen können. Ein übersichtlich gestaltetes graphisches Interface soll dem Benutzer gestatten, schnell zu erkennen, ob eine vorgegebene Aminosäure-Sequenz mit den theoretisch berechneten Massenlisten hinreichend übereinstimmt, so dass man von einer Identifizierung sprechen kann.
Aufgabe des Programms ist es, die Massen einer Peptidsequenz in einem experimentellen MS-Spektrum zu identifizieren, um so in der Lage zu sein, die Position einer möglichen co- oder posttranslationalen Modifikation zu erkennen. Dabei wurde, wie in der ganzen Arbeit, darauf geachtet, dass solche Parameter mit Hilfe von Pulldown-Menüs eingegeben werden können, um zum einen zu verhindern, dass durch Fehleingaben falsche Ergebnisse erzeugt werden oder durch unsinnige Eingaben keine Berechnungen durch das Programm erfolgen können.

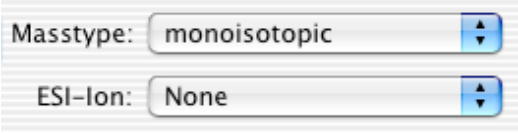


Abbildung 22: Spektrum-relevante Menüs des Programms *FINDYSERIES*

4.9.2 Umsetzung

Da für *FINDYSERIES* eine effiziente und schnelle Interaktion mit dem Benutzer gefordert ist, wurde hier auf die Benutzung eines Web-Browsers zur Darstellung der

Ergebnisse verzichtet. Der notwendige Datenverkehr mit dem Internet wurde hier mittels XML-Containern, die die Daten übertragen, erreicht.

4.9.3 Das Benutzer-Interface

Das Benutzer-Interface wurde zweigeteilt in einen Eingabe-Bereich und einen großen Ausgabe-Bereich.

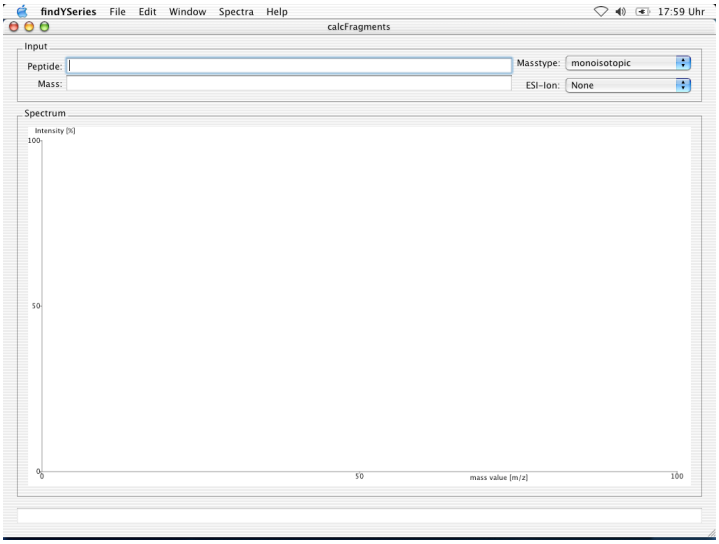


Abbildung 23: Benutzer-Interface des Programms *FINDYSERIES*

4.9.4 Manuelle Zuordnung der Massen zu Fragment-Ionen

Als erstes wurde eine Möglichkeit implementiert, die es dem Benutzer gestattet, durch manuelle Eingabe von Sequenzen die tatsächliche Sequenz einem experimentellen Spektrum zuzuordnen.
Dazu muss der Benutzer das verwendete ESI-Ion mit Hilfe des entsprechenden Popup-Menüs einstellen. Die Eingabe der Sequenz erfolgt manuell in das Eingabefeld <Peptide>. Das Einlesen einer Liste von Massen geschieht über den Menüpunkt <New>. Als Datenformate werden hier das Sequest dta-Format und das Micromass pkl-Format unterstützt. Die Massen des *in silico* berechneten Spektrums werden mit dem eingelesenen, experimentellen Spektrum verglichen.
Die Ausgabe erfolgt im unteren Bereich des Interfaces. Nach jeder Änderung im Eingabefeld <Peptid> wird verglichen, welche Massen des theoretischen Spektrums vorhanden sind, und die Treffer werden rot dargestellt.
Dieser Modus dient ausschließlich einer manuellen Überprüfung der Übereinstimmung von *in silico* berechneten und einem experimentellen Spektrum. Der Ansatz ist nicht geeignet für eine vollautomatische Identifizierung eines Proteins oder für seine de-novo Sequenzierung. In der täglichen Arbeit hat sich dieses Werkzeug jedoch als sehr nützlich erwiesen, da es erlaubt, schnell und gezielt bestimmte Sequenzen zu überprüfen und unsinnige Sequenzen schnell auszuschließen. Analog wird zu der visuellen Auswertung auch noch ein Bewertungsreport erzeugt, der angibt, wie viele der erwarteten Massen im Massenspektrum wieder gefunden werden.

4.9.5 Automatische Sequenzierung von Proteinen

Der rasante technische Fortschritt bei der Entwicklung von Massenspektrometern hat u.a. auch zu einer bedeutenden Verbesserung der Auflösung der detektierbaren Massendifferenzen geführt. So ist es heute möglich, die Massen der Fragment-Ionen

mit einer Genauigkeit von mindestens 0,02Da aufzunehmen. Das heißt: Es ist ohne Probleme möglich, die Aminosäuren wie in Tabelle 8 aufgelistet zu identifizieren.

Aminosäure	Abkürzung	Lettercode	Monoisotopische Masse	Differenz	Anmerkung
Glycin	Gly	G	57,021		
Alanin	Ala	A	71,037	14,016	
Serin	Ser	S	87,032	15,995	
Prolin	Pro	P	97,053	10,021	
Valin	Val	V	99,068	2,015	
Threonin	Thr	T	101,048	1,980	
Cystein	Cys	C	103,009	1,961	
Isoleucin	Ile	I	113,084	10,075	
Leucin	Leu	L	113,084	0,000	Kein Unterscheidung mit Isoleucin möglich
Asparagin	Asn	N	114,043	0,959	
Asparaginsäure	Asp	D	115,027	0,984	
Glutaminsäure	Glu	E	128,059	13,032	
Lysin	Lys	K	128,095	0,036	
Glutamin	Gln	Q	129,043	0,948	
Methionin	Met	M	131,041	1,998	
Histidin	His	H	137,059	6,018	
Phenylalanin	Phe	F	147,068	10,009	
Arginin	Arg	R	156,101	9,033	
Tyrosin	Tyr	Y	163,063	6,962	
Tryptophan	Trp	W	186,079	23,016	

Tabelle 8: monoisotopische Massendifferenzen der Aminosäuren

Außer bei den in der Masse identischen Aminosäuren Leucin und Isoleucin lassen sich also alle Aminosäurenfragmente auf Grund der Massendifferenz unterscheiden. Wegen ihrer Massengleichheit werden Leucin und Isoleucin mit einem ‚X' bezeichnet und nicht wie sonst üblich mit ‚I' oder ‚L'.

4.9.6 Integration der Bestimmung der Sequenz mit Hilfe der Software *MASCOT*

Wie oben schon erläutert, stellt die computerbasierte Spektreninterpretation mit dem Programm *MASCOT* von MS/MS-Spektren von Peptiden gegenwärtig die am weitesten fortgeschrittene Technologie zur automatischen Identifizierung von Proteinen dar. Im DKFZ steht eine Lizenz des Softwarepaketes zur Verfügung. Leider ist im Moment ein Zugriff auf die *MASCOT*-Software nur über ein Web-Interface möglich. Es entspricht natürlich keiner guten Benutzerführung, den Benutzer erst mittels des *MASCOT*-Webinterfaces die Peptidsequenz ermitteln zu lassen und diese anschließend in das entsprechende Eingabefeld <Peptide> des Programms *FINDYSERIES* eingeben zu lassen. Es musste daher eine Möglichkeit entwickelt werden, die Daten direkt aus der Ausgabe des Webservers in das Programm einzulesen. Da der Prozess einer direkten Einbindung von Daten, die mittels einer externen Anwendung aus dem Internet extrahiert werden, in ein lokal ausgeführtes Programm noch nicht als eine Standardprozedur betrachtet werden kann, war die Entwicklung einer Reihe von neuen Werkzeugen notwendig, um

diesen Zugang zu realisieren. Aus diesem Grunde soll im folgenden Einschub auf die technische Realisierung der Interaktion näher eingegangen werden.

Bei einem Webserver handelt es sich um ein Programm, das über den Port 80 mittels des HTT-Protokolls mit anderen Rechnern in Verbindung steht. Man braucht daher lediglich eine Webanfrage zu simulieren, die genau wie die eines Browsers aussieht. Eine Analyse des entsprechenden Formulars, das von der Firma Matrixscience[1] entwickelt worden ist, ergab, dass die Anfragen mittels eines POST-Request nach dem Hypertext Transfer Protocol (HTTP) erfolgen.
Man geht daher wie folgt vor: Das Betriebssystem UNIX stellt Funktionen bereit, die eine Low-Level Kommunikation über das TCP/IP-Protokoll mit einem anderen Rechner gestatten. Mit diesen Funktionen wird nun die Anfrage an den Webserver gesendet, der wiederum die *MASCOT*-Software aufruft und als Antwort eine normale HTML-Datei zurückliefert. In dieser HTML-Antwort befindet sich nun die URL, unter der das eigentliche Ergebnis abgerufen werden kann. Dasselbe Vorgehen wird auch von der Firma *MASCOT* für Ihren <Mascot-Daemon> verwendet. Allerdings ist hier keine routinemäßige Analyse der Daten in der gelieferten Datei vorgesehen.
Die *MASCOT*-Software legt das Analyseergebnis in Form einer einfachen, jederzeit lesbaren Datei (Flat-File) auf dem Webserver an. In diesem File sind die ermittelten Werte dokumentiert. So sind auch alle Treffer und Sequenzen darin enthalten, die bei dem Vergleich von gemessenen und berechneten Spektren ermittelt wurden. Diese Ergebnisdokumentation soll an zwei Zeilen der Ausgabedatei exemplarisch dargestellt werden.

.

h1_text=DYNAMIN LIKE PROTEIN.- Dictyostelium discoideum (Slime mold).
h1_q1=0,1148.549530,-0.017316,215,225,6.00,GTDAMEVLTGR,9,0000000000000,
60.58,1,001100200000000,0,0,1600.600000

Listing 1: Struktur des *MASCOT*-Ergebnis-Files

Die erste Zeile enthält die genaue Proteinbezeichnung. Die zweite Zeile entspricht dem ersten Treffer der Anfrage. Dazu werden durch Kommata getrennt die theoretische Masse der Sequenz, die Differenz zur gemessenen Masse und die Sequenz sowie den Score, den *MASCOT* intern errechnet hat, mit angegeben. Die Bedeutung der anderen Werte konnte leider nicht ermittelt werden, sie hatten aber für das weitere Vorgehen keine Bedeutung. Diese Datei kann nun relativ einfach ausgewertet werden. Die dazu nötigen Routinen konnten schnell entwickelt werden. Leider ist vom Hersteller keine Möglichkeit vorgesehen, dieses Ergebnis-File durch eine einfache HTTP-Anfrage zu erhalten und auszuwerten. Es musste daher ein kleines CGI-Skript entwickelt werden, das auf dem *MASCOT*-Webserver installiert wird und diese Ergebnisdatei zurückliefert. Die zurück gelieferte Datei enthält alle benötigten Daten, die von der *MASCOT*-Software erzeugt worden sind, und die eigentliche Auswertung kann durch das Programm *FINDYSERIES* erfolgen.
Dieser Vorgang wird durch Auswählen des Menüpunktes <Query Mascot> ausgelöst, und das Ergebnis ist je nach gewählter Hardware in der Zeit zwischen 10 Sekunden (Server mit parallelen Pentium III Prozessoren) und bis zu sechs Minuten auf einer Einprozessor-Maschine (hier eine schon etwas älteren Sun-Server mit R1000 Prozessor) zurückgeliefert.

[1] http://www.matrixscience.com/cgi/index.pl?page=/

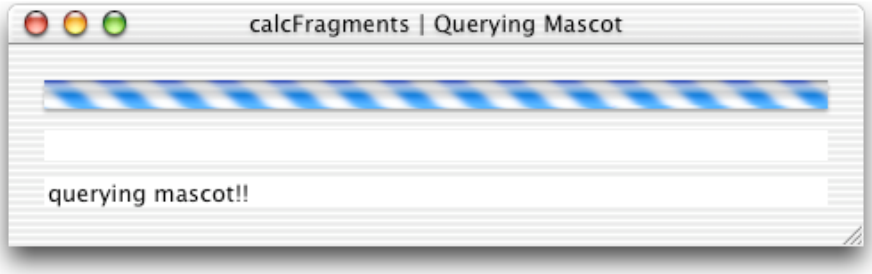


Abbildung 24: Interaktion mit dem *MASCOT*-Server

Die eigentliche Auswertung der bereitgestellten Ergebnisse durch das Programm *FINDYSERIES* erfolgt innerhalb von wenigen Millisekunden. Es wird nur das beste Ergebnis der *MASCOT*-Anfrage in Betracht gezogen. Die Bewertungsfunktion liefert den Logarithmus der Wahrscheinlichkeit P zurück. Die Wahrscheinlichkeit P gibt an, ob das Suchergebnis nur auf einer zufälligen Übereinstimmung einzelner Massen basiert oder ob eine echte Übereinstimmung der Messwerte mit den berechneten Bruchstücken besteht.

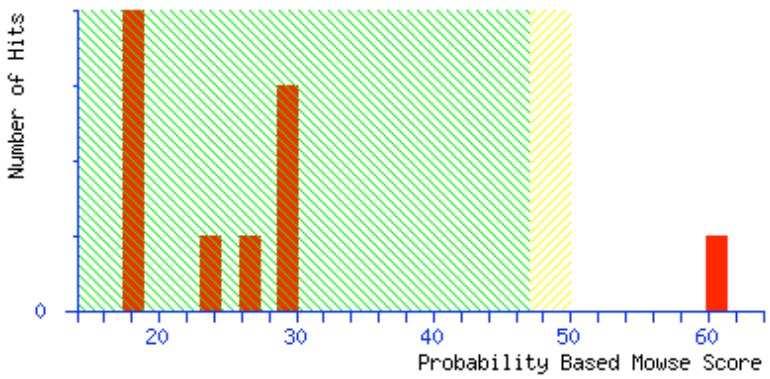


Abbildung 25: zurückgelieferte Wahrscheinlichkeitswerte

Sollte der Score, den *MASCOT* zurückliefert, kleiner als 50 sein, so wird auf das Ergebnis verzichtet, da dieser Wahrscheinlichkeitswert zu sehr einem zufälligen Ereignis entspricht. Ergebnisse, die jedoch einen höheren Wahrscheinlichkeitswert besitzen, können als richtig zugeordnet betrachtet werden.

4.9.7 Beispiel

Die Bestimmung der Position von einer post-translationalen Acetylierung soll hier nun beispielhaft durchgeführt werden. Dabei wird auf Daten von Professor Lehmann zurückgegriffen[44], der ein Experte für die Bestimmung von Modifikationen an Proteinen und deren Analyse mit Hilfe der MS ist. Für eine anschauliche Darstellung des Prinzips wurde auf eine Acetylierung zurückgegriffen, da es hier nur zu einer einzigen Massenverschiebung kommen kann, und nicht wie bei Glykosylierungen das Protein noch mit unterschiedlichen Glykanen dekoriert sein kann.

4.9.7.1 Experimentelle Vorbereitung

Als Beispiel wurde das schon in der Einleitung diskutierte Protein ‚Dynamin like Protein from Dictyostelium discoideum'[44] verwendet. Es wurde einem Verdau durch das Enzym Trypsin unterzogen. Dabei entstanden 85 verschiedene Bruchstücke. Nachdem diese Bruchstücke aufgereinigt wurden, konnten davon MS/MS-Spektren

aufgenommen und ihre Sequenz mit *MASCOT* bestimmt werden. Diese interpretierten MS/MS-Spektren werden nun mit Hilfe des Programms *FINDYSERIES* auf mögliche Acetylierungen untersucht.
Dazu wird als erstes das zu untersuchende MS/MS-Spektrum des Peptids geladen. Dies geschieht sehr leicht mit Hilfe des standardisierten <Datei Öffnen>-Dialogs des Betriebssystems.

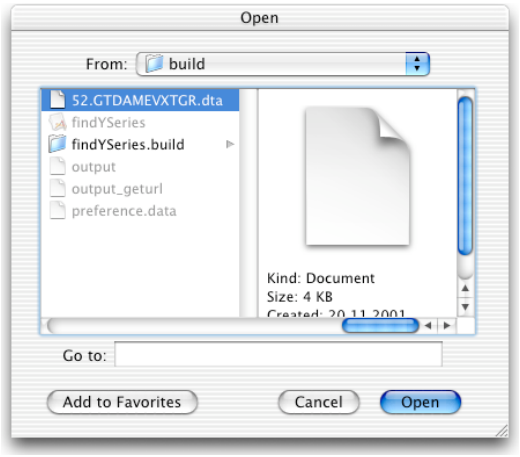


Abbildung 26: Öffnen-Dialog des Programms *FINDYSERIES*

Nachdem das Massenspektrum geladen worden ist, wird es sofort im unteren Bereich des Benutzerinterfaces dargestellt.
Bei der Darstellung der Spektren wird die Intensität der einzelnen Massen immer in Prozent angegeben. Dies führt zu der folgenden Darstellung:

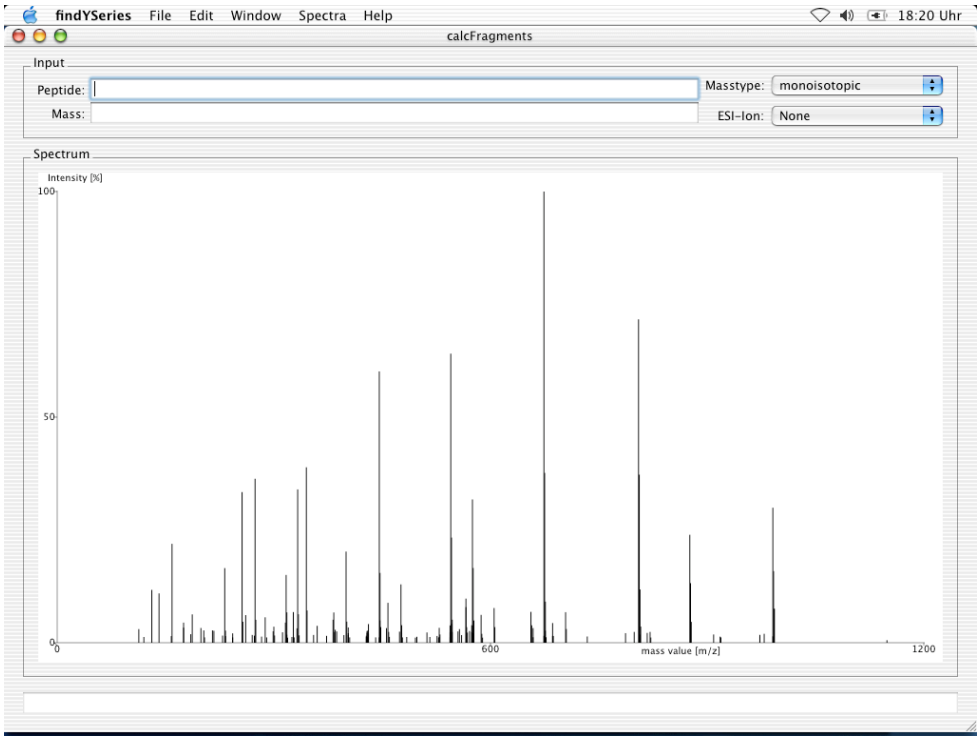


Abbildung 27: Darstellung des eingelesenen Spektrums

Nun kann begonnen werden, die Sequenz–Informationen aus dem Massenspektrum zu ermitteln. Dazu müssen die wichtigsten Messbedingungen dem Programm mitgeteilt werden. Dieses sind die verwendete Signal-Art und das verwendete Esi-

Ion. Sind diese beiden Parameter den Messbedingungen entsprechend eingegeben worden, so kann auf manuelle Art und durch Benutzung der *MASCOT*-Software versucht werden, die Sequenz zu bestimmen. Bei der manuellen Analyse können verschiedene Sequenzen eingegeben und deren berechnete Fragment-Ionen mit den experimentellen Massen verglichen werden. Dieses geschieht in dem Eingabefeld <Peptide> des Userinterfaces.

Abbildung 28: Eingabe der Sequenzinformation und Darstellung der berechneten Masse

Aus der eingegebenen Sequenz wird aufgrund der bekannten Fragmentierungsmuster sofort ein theoretisches Spektrum berechnet, das mit den gemessenen Massen verglichen wird. Übereinstimmende Massen werden in rot dargestellt.

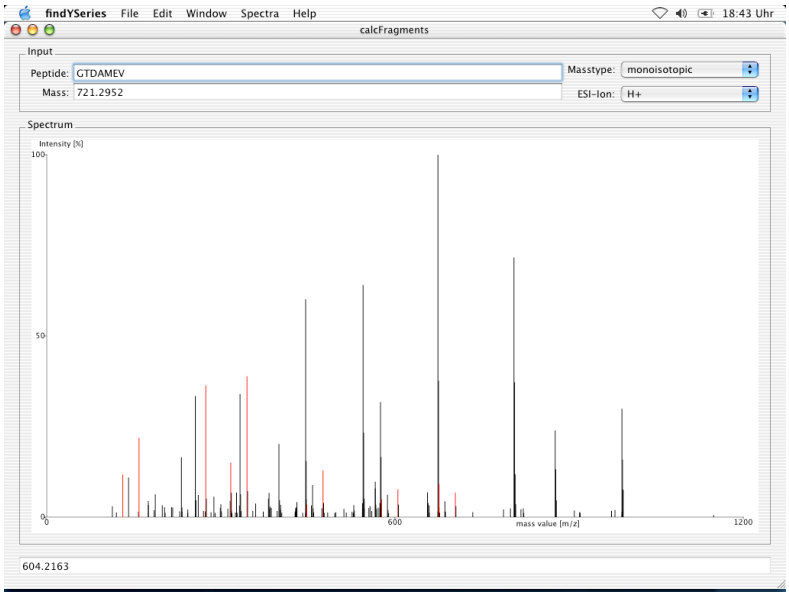


Abbildung 29: Übereinstimmende Massen in gemessenem und theoretischen Spektrum

Durch Anklicken mit der Maus kann dann die Bezeichnung eines Ions (siehe Abschnitt ‚Tryptischer Verdau eines Proteins') angezeigt werden, das sich aus seiner Masse im in silico Spektrum ergibt. Dabei kann sehr leicht an Hand des generierten Bewertungsreports festgestellt werden, ob es sich um Serien von Fragmenten handelt. So können in dem obigen Beispiel bis auf den ‚y-10' Peak alle Y-Fragmente gefunden werden.
Danach kann begonnen werden, nach co- und posttranslationalen Modifikationen des Peptids zu suchen. Dies geschieht folgendermaßen: Nacheinander können über das Submenü <Modifications> aus dem <Spectra>-Menü alle vom Programm unterstützten Modifikationen ausgewählt werden.

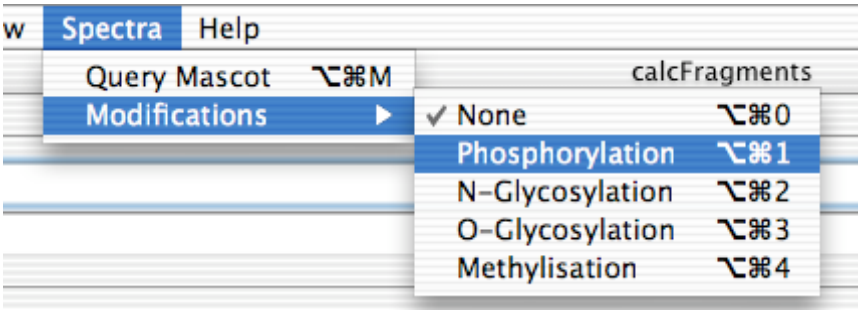


Abbildung 30: Menü zur Suche nach Modifikationen

Damit kann man nun nacheinander jede dieser Möglichkeiten auswählen. Je nach ausgewähltem Menüpunkt werden nun in das theoretische Spektrum Peaks hinzugefügt, die zum Beispiel einem y-5-Ion entsprechen, dessen Massenwert um die Masse einer Phosphorgruppe oder eines N-Acetyl-Glucosamin erhöht ist. Dieses theoretische Spektrum wird nun mit dem gemessenen Spektrum verglichen. Dabei werden die Massen, die einem modifizierten Fragment entsprechen könnten, grün eingefärbt.

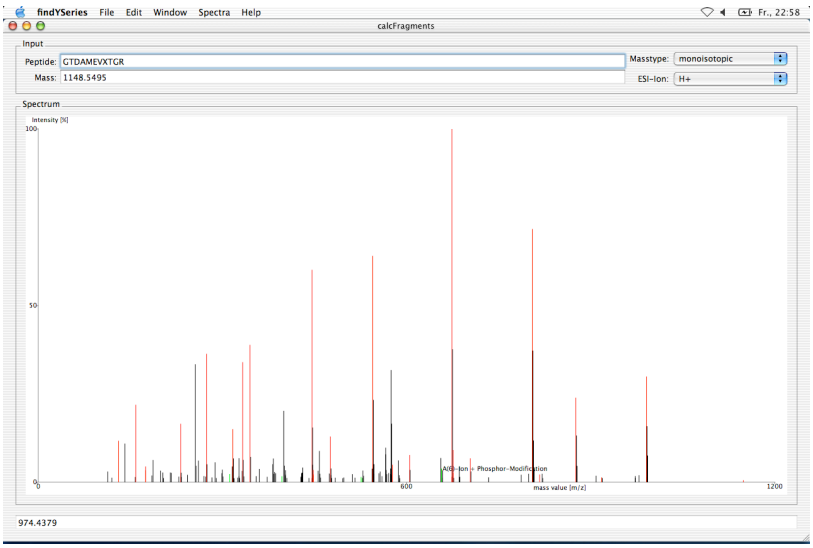


Abbildung 31: Spektrum mit allen zugeordneten Peaks

Auf diese Art und Weise ist eine sehr schnelle visuelle Kontrolle des Spektrums möglich, und man erhält sehr schnell einen Hinweis darauf, ob man weitere Spektren von diesem Peptid-Fragment aufnehmen soll.

4.9.7.2 Erzeugung eines Reports

Es ist dem Benutzer natürlich nicht zuzumuten, für jede neu eingegebene Sequenz das Ergebnis durch Anklicken jeder Masse mit der Maus zu erfahren, ob sie auch mit dem *in silico* generierten Spektrum übereinstimmt. So war es nötig eine Bewertungsfunktion zu entwickeln, die es dem Benutzer gestattet, auf einen Blick zu erfahren, welche und wie viele übereinstimmende Massen in dem experimentellen Spektrum und den sich aus der Sequenz ergebenden theoretischen Spektrum enthalten sind. Dies sei an dem folgenden Beispiel verdeutlicht. Es wird ein Spektrum mittels des Menüpunktes <New> geladen und die Sequenz eingetragen:

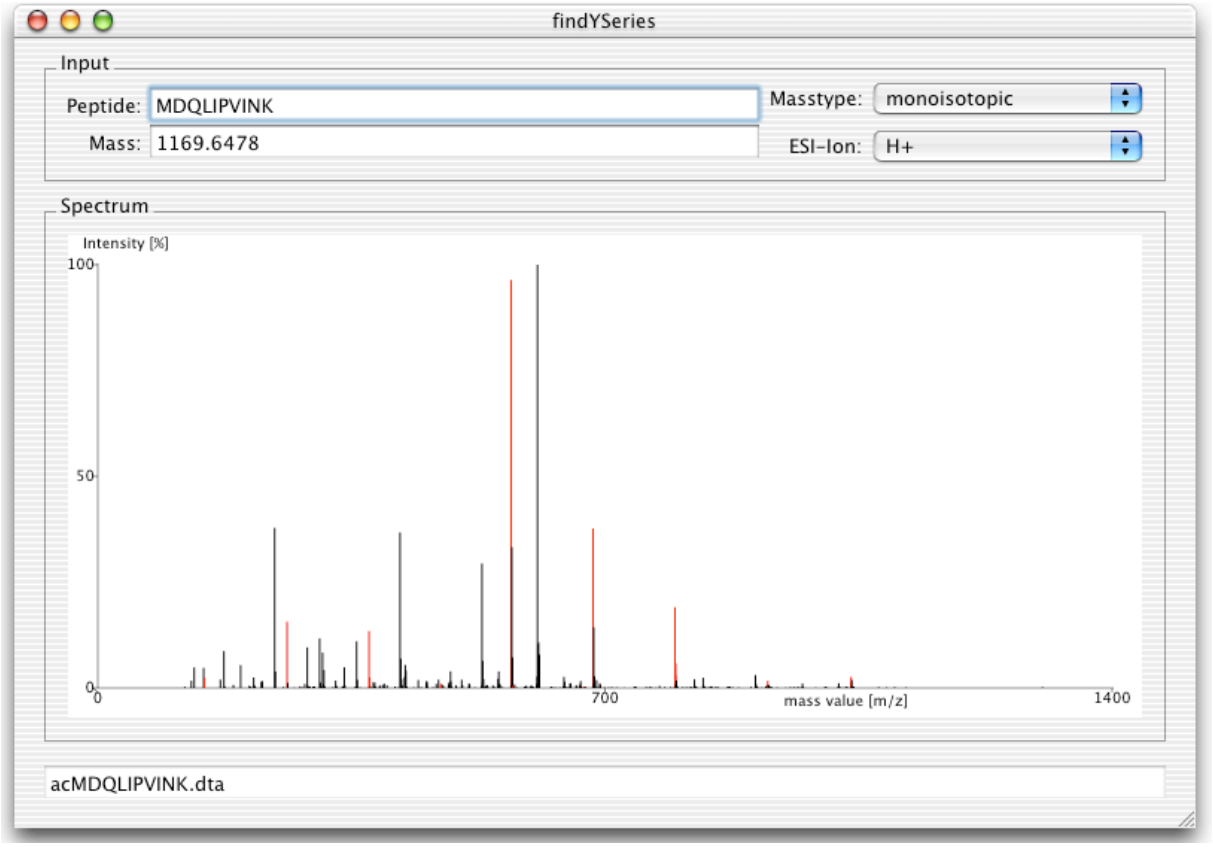


Abbildung 32: Spektrum der Sequenz MDQLIPVINK

Dann werden das Esi-Ion, die Art der ermittelten Massen und die gesuchte Modifikation mit Hilfe der Menüs eingestellt. Abhängig von diesen Eingaben wird jetzt ein Report generiert.

File: acMDQLIPVINK.dta					
#	a-Ion	b-Ion	Sequence	y-Ion	#
1	+	+	M	*	10
2	+	+	D	*	9
3	+	* +	Q	*	8
4	+	+	L	*	7
5	*	+	I	*	6
6	*	* +	P	* +	5
7		*	V	*	4
8	+		I	*	3
9	+		N	*	2
10			K		1
I found 27 peak(s) from 60 possible					
* means peak is present, + peak is modified with Acetyl					

Tabelle 9: Report mit Bewertung der gefundenen Peaks

Der Report zeigt sehr schön, dass das untersuchte Fragment die Acetylierung am N-Terminalen Ende trägt. So sind außer dem sehr schwer zu messenden acetylierten b-10-Ion und dem nicht entstehenden b-7-Ion alle acetylierten a- oder b-Ionen der Sequenz enthalten. Gleichzeitig sind fast keine acetylierten y-Ionen zu sehen. Mit dieser Methode konnten nacheinander 153 von Prof.-Lehmann zur Verfügung gestellten MS/MS-Spektren analysiert werden. Es wurden dabei keine weiteren Modifikationen gefundenen. Für die komplette Analyse wurden weniger als zwei Stunden benötigt.

4.10 Entwicklung von Hilfswerkzeugen

Bei der Entwicklung des Programms *FINDYSERIES* wurde sehr schnell klar, dass es dringend nötig war, weitere einfache Hilfsklassen zu entwickeln, die den Benutzern auch in anderen Programmen zur Verfügung stehen können. Bei der Fehlersuche und dem Austesten der Programme mussten immer wieder dieselben Berechnungen für einzelne Peaks oder eine Substanz gemacht werden, um zu Überprüfen, ob ein Peak richtig zugeordnet ist oder nicht. Dieses legt die Vermutung nahe, dass auch bei der direkten Auswertung von Massenspektren diese Berechnungen schnell und einfach vorgenommen werden müssen. Außerdem war es nötig, die in der Arbeit geforderte webbasierte Bereitstellung von Daten auch für dieses Projekt zu untersuchen. So sollte beispielhaft untersucht werden, inwieweit es möglich ist, eine Online-Hilfe für das Programm *FINDYSERIES* zur Verfügung zu stellen, die tagesaktuell ist, aber nicht von einer Anbindung an das Internet abhängig sein soll.

4.10.1 Konvertierung von Massenwerten in Ionen

In einem Massenspektrum werden immer Intensitäten gegen m/z (Masse/Ladung) aufgetragen. Dies führt dazu, dass doppelt geladene Ionen auch nur bei der halben Masse im Spektrum gemessen werden. Es war daher häufig notwendig zu überprüfen, ob ein Fragment mit unterschiedlichen Ladungen in den Spektren vorkommt. Dieses machte es nötig, zwei weitere Klassen zu entwickeln, die diese

Aufgabe schnell und einfach lösen können. Zum einen ist es die Klasse zum Berechnen der Masse des Ions:

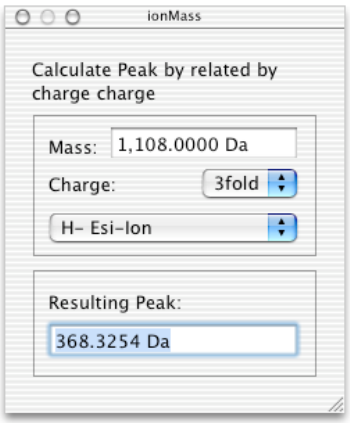


Abbildung 33: Berechnung des Ions aus einer Masse

Zum anderen die Klasse zum Berechnen des Massenwerts aus dem Ion.

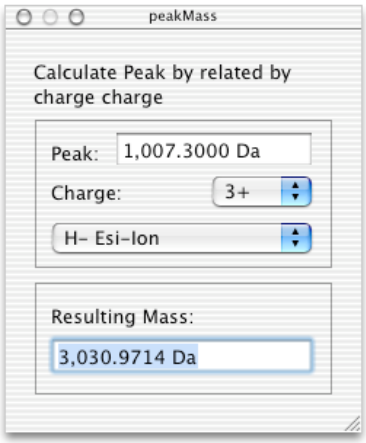


Abbildung 34: Berechnung der Masse aus einem Ion

4.10.2 Quervernetzung der Applikation mit der *SWISSPROT*-Datenbank

Ein sehr wichtiger Aspekt für die weitere Entwicklung von Anwendungen im Bereich der Glykobiologie besteht in der Quervernetzung von Datenbanken und den einzelnen Applikationen. So soll gewährleistet sein, dass Daten aus verschiedenen Quellen einfach vom Anwendungsprogramm genutzt werden können. Dies soll hier anhand der Verknüpfung mit der *SWISSPROT*-Datenbank gezeigt werden. Da es sich um eine Datenbank handelt, die allgemein im Internet verfügbar ist und zentral gepflegt wird, konnte hier nicht auf die Technik eines XML-Containers zurückgegriffen werden, der die Daten von der Datenbank an die lokalen Anwendungen übermittelt. Stattdessen wurde hier der Umweg über den Webserver gewählt. Es wird also einfach die HTML-Antwort auf eine Anfrage an den Webserver ausgewertet. Dieses geschieht wie folgt: Man gibt in den oberen Teil des Fensters entweder den Entryname eines in der *SWISSPROT* enthaltenen Proteins, die Cross-Reference Bezeichnung oder die Primary Accession Number ein und klickt auf den Button <Query Swissprot>. Das Programm setzt daraufhin die URL der Anfrage zusammen und übermittelt sie an den *SWISSPROT*-Webserver. Die zurückgeschickte HTML-Datei wird lokal ausgewertet und die Ergebnisse werden in die entsprechenden Felder der lokalen Anwendung eingetragen.

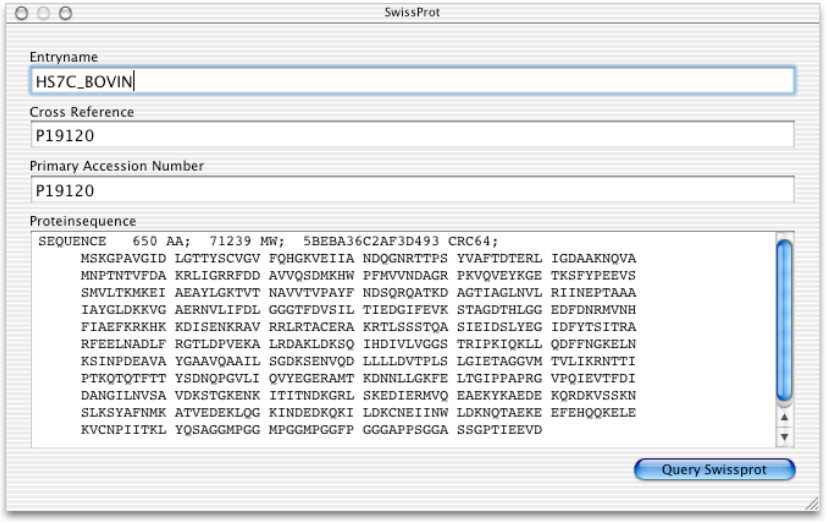


Abbildung 35: Ergebnis einer Anfrage an die *SWISSPROT*

Es ist so ohne größeren Aufwand möglich, die Sequenz des interessierenden Proteins zu erhalten, und sie steht nun für weitere Berechnungen oder Auswertungen zur Verfügung.

4.10.3 Darstellung einer webbasierten Hilfe

Ein großes Problem bei der Erstellung von Applikationen, die nicht webbasiert sind, ist es, die einzelnen Komponenten und Daten immer auf einem aktuellen Stand zu halten. Dieses soll hier exemplarisch an der Anwendung *MAINHELP* gezeigt werden. Bevor eine Hilfeseite der lokalen Anwendung angezeigt wird, wird eine Web-Anfrage gestartet, die überprüft, ob eine aktuellere Version der Hilfen auf dem Webserver bereitgestellt wird. Ist dem so, so wird die aktuellere Hilfeseite geladen und ersetzt die gespeicherten Daten. Dazu wird eine Anfrage in der folgenden Form an den Webserver gestellt:

```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE mainHelpQuery (View Source for full doctype...)>
<mainHelpQuery>
<query>
<helptopic>Selecting Masstype</helptopic>
</query>
</mainHelpQuery>
```

Listing 2: XML-Container der Anfrage

Der Webserver liefert daraufhin einen XML-Container mit der korrespondierenden Hilfeseite zurück, in dem der Dateiname und das Datum der letzten Änderung codiert sind.

```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE mainHelpAnswer (View Source for full doctype...)>
<mainHelpAnswer>
<answer>
<helptext>
```

```
<file>http://www.dkfz.de/spec/findYSeries/helpfiles/selected_masstype_page_1.rtfd</
file>
<last_change>
<day>13</day><month>09</month><year>2003</year>
</last_change>
</helptext>
</answer>
</mainHelpAnswer>
```

Listing 3: XML-Container der Hilfeseite

Dieser XML-Container wird nun ausgewertet und falls die Seite aktueller ist, erfolgt das Herunterladen der Hilfeseite und ersetzt fortan die alte Hilfeseite. Auf diese Art und Weise ist auch die Aktualität gewährleistet, wenn momentan keine Anbindung an das Internet besteht. Außerdem wird so auch die Belastung des Internets so gering wie möglich gehalten. Da die Hilfeseiten im Allgemeinen nur eine geringe Dateigröße besitzen, bemerkt der Benutzer von diesem Vorgang nichts. Die Darstellung der Hilfeseite erfolgt wieder mit der eigenen Applikation. Dieses erfolgt in der Form eines Hilfecenters, wie es für heutige Programme Standard ist. So kann der Benutzer im oberen Teil des Centers nach dem passenden Begriff suchen. Nach Auswahl des Begriffes erhält er die entsprechenden Informationen angezeigt.

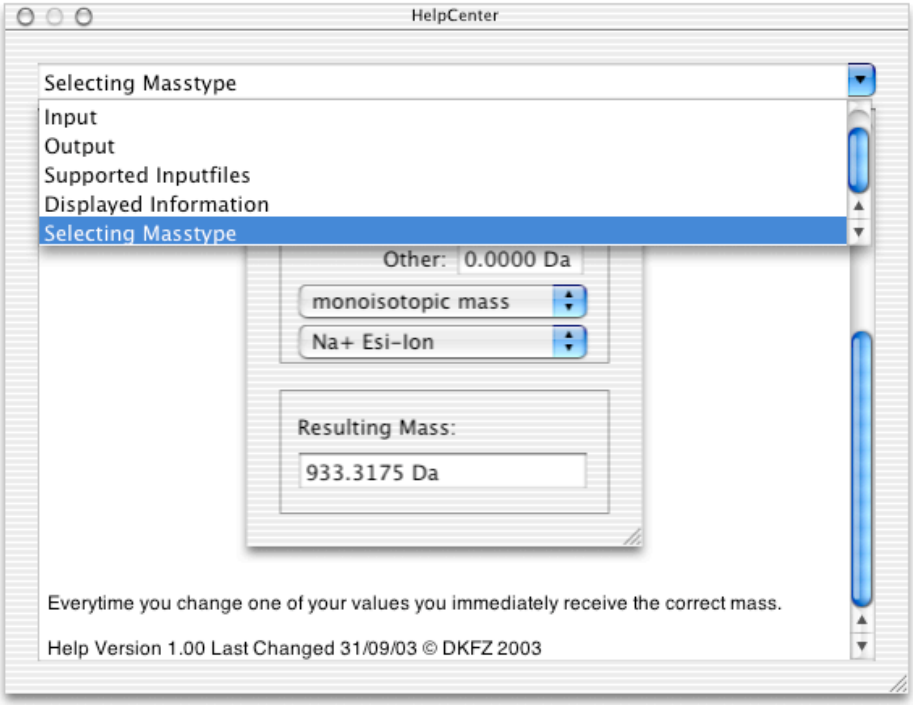


Abbildung 36: Darstellung des HelpCenters

Man hat so die Möglichkeit die Vorteile einer lokal ausgeführten Applikation, wie Geschwindigkeit und gute und schnelle Interaktion mit den Vorteilen einer webbasierten Anwendung, wie gute Aktualität und zentrale Beseitigung eventueller Fehler, miteinander zu verbinden.

4.11 Ergebnis und Diskussion

Mit dem hier vorgestellten Programm ist es schnell und einfach möglich, in dem MS- oder MS/MS-Spektrum einer Peptidsequenz nach co- und posttranslationalen

Modifikationen zu suchen. Mit der Anbindung an einen *MASCOT*-Server konnte gezeigt werden, dass über definierte Schnittstellen auch das Einbinden eines bereits bestehenden Softwarepaketes in eine Applikation möglich ist. Durch Verzicht auf eine browserbasierte Darstellung der Ergebnisse und Handhabung der Benutzereingaben wurde eine bedeutend bessere Interaktion mit dem Benutzer möglich. Allerdings erkauft man diesen Vorteil durch eine größere Betriebssystem- und Hardwareabhängigkeit. Dafür muss sich der Benutzer aber nicht mehr um unterschiedliche Implementierungen von Javascript oder Cascading Style Sheets in den Browsern von unterschiedlichen Herstellern Gedanken machen.
Für die Zukunft ist es natürlich zwingend nötig die automatischen Analysemöglichkeiten des Programms zu verbessern. So sollte es möglich sein, die Massen, die zu einer Serie von Fragmenten gehören, unterschiedlich anzufärben. Dadurch wird gerade bei der manuellen Auswertung des Spektrums eine visuelle Kontrolle des Ergebnisses erleichtert.

4.12 Entwicklung von Algorithmen zur Berechnung der Massen von Fragmenten und Ionen von Glykanen.

Nachdem die Sequenz eines glykosylierten Peptids und auch die Position der Modifikation gefunden ist, gilt es im Folgenden, die Sequenz und Komposition des Kohlenhydrats zu ermitteln. Wie schon in der Einleitung beschrieben, werden auch hierfür Methoden der MS intensiv genutzt. Zur Unterstützung der automatischen Auswertung und Interpretation von MS-Spektren haben wir die Entwicklung von Algorithmen vorangetrieben, die es ermöglichen, die Berechnungen der Massen von Fragmenten und Fragment-Ionen von Glykanen automatisch vorzunehmen. Primäres Ziel der Entwicklung war eine einfache, möglichst automatische Zuordnung der Massen in einem gemessenen Spektrum zu ermöglichen. Der hier entwickelte Algorithmus und seine Implementierung werden im Folgenden mit dem Namen *GLYCO-FRAGMENT* gekennzeichnet.

4.12.1 Grundlagen des *GLYCO-FRAGMENT*-Algorithmus: Berechnung von Fragmenten

Grundlage des *GLYCO-FRAGMENT* Algorithmus bildet die Berechnung aller Fragmente in die Glykane zerfallen können. Die Bezeichnung der entstehenden Fragmente geht auf eine Arbeit von Domon und Costello[58] zurück, die inzwischen allgemeinen Eingang in die Literatur gefunden hat. Der implementierte *GLYCO-FRAGMENT*-Algorithmus benötigt als Eingabe eine komplette topologische Beschreibung des zu interpretierenden Glykans. Dazu wird zuerst mit Hilfe des Programms *SWEET-II*[68, 69] eine komplette chemische Struktur erzeugt, die alle Atome des Kohlenhydrats enthält. *SWEET-II* ist in der Lage, die in dieser Arbeit verwendete erweiterte IUPAC-Nomenklatur für Oligosaccharide zu verarbeiten. Der *SWEET-II* Algorithmus analysiert die Struktureingabe und erkennt die enthaltenen Monosaccharide sowie deren Verknüpfungen und Verzweigungen. Basierend auf einer Datenbank, die komplette Strukturen von Monosacchariden enthält, setzt *SWEET-II* die einzelnen Bausteine entsprechend den Angaben über Verknüpfung und Verzweigung zusammen, so dass eine komplette topologische Beschreibung des gewünschten Glykans entsteht.
Zurzeit enthält die von *SWEET-II* benutzte Datenbank etwa 450 molekulare Bausteine. Diese können unter Angabe der Position mit 20 verschiedenen oft vorkommenden Substituenten substituiert werden. Eine Liste dieser Substituenten

befindet sich im Anhang der Arbeit. Sowohl die Liste der verfügbaren Monosaccharideinheiten als auch der Substituenten wurde im Laufe dieser Arbeit ständig erweitert. Als Eingabe werden lediglich die räumlichen Strukturen der Bausteine benötigt, die mit allen üblichen Programmen zur Konstruktion von Molekülen erzeugt werden können. Allerdings muss die Benennung der Atome, an denen eine Substitution erfolgen soll, der allgemein üblichen Nomenklatur entsprechen. Mit diesem Repertoire an Bausteinen können ohne größere Schwierigkeiten auch sehr komplexe Verbindungen, die von wissenschaftlichem Interesse sind, mittels *SWEET-II* erzeugt werden.
Nachfolgend benutzt der *GLYCO-FRAGMENT* Algorithmus die komplette topologische Information sowie die ebenfalls von *SWEET-II* bereitgestellten Linkage- und Atominformationen, alle möglichen A-, B-, C-, X-, Y- und Z-Fragmente entsprechend der Definition von Domon und Costello[58] zu berechnen. Dieses geschieht wie folgt: als erstes wird mit Hilfe der Verknüpfungsinformation ein Baum aufgebaut, der der Topologie des Zuckers entspricht. Durch Spalten aller glykosidischen Bindungen werden zunächst nacheinander alle B-, C-, sowie X-, Y-Fragmente berechnet. Anschließend werden durch das Spalten von jeweils zwei Ringbindungen alle X- und A-Fragmente berechnet. Im letzten Schritt wird die Masse des ladungtragenden Atoms zu der Masse des Fragmentes addiert und man erhält so eine Liste aller Fragment-Ionen.
Die Erzeugung einer kompletten topologischen Beschreibung der zu analysierenden Verbindung, die alle Atome explizit enthält, hat den Vorteil, dass der beschriebene Algorithmus ohne Änderungen für alle Arten von Veränderungen der chemischen Struktur angewendet werden kann. Für den *GLYCO-FRAGMENT* Algorithmus ist es unerheblich, welche Substitution an welcher Position vorgenommen wurde. Er spaltet lediglich die vorgegebene Struktur an den entsprechenden Stellen und summiert dann die Massen der Atome des rechten und linken Fragments. Alternative Algorithmen, die auf den abgelegten Massen für die entsprechenden Bausteine aufbauen (*GLYCOMOD*, *GLYCOMASS*), würden für jede Art einer Substitution und deren Position veränderte Listen von Massen benötigen.
Der die Geschwindigkeit des Algorithmus bestimmende Schritt ist die Erzeugung der kompletten Topologie des Glykans mittels des *SWEET-II* Algorithmus. Allerdings ist auch diese Berechnung so schnell, dass für die interaktive Anwendung des Algorithmus auf einzelnen Strukturen über das Web-Interface für den Benutzer kaum merkbare Verzögerungen entstehen. Sollen aber die Fragmente von einigen 10.000 bis 100.000 Strukturen aus einer Datenbank (siehe *GLYCO-SEARCH-MS*) berechnet werden, um sie mit experimentellen Spektren zu vergleichen, so wird es notwendig sein, den *SWEET-II* Algorithmus im Hinblick für die vom *GLYCO-FRAGMENT* Algorithmus benötigten Informationen zu optimieren.

4.12.2 Anforderungen an die Eingabe für den *GLYCO-FRAGMENT* Algorithmus

Damit die Interpretation der gemessenen Spektren mit Hilfe von *GLYCO-FRAGMENT* tatsächlich Eingang in die tägliche Praxis von Glykowissenschaftlern findet, ist es wichtig, dass möglichst viele Ansätze und Methoden unterstützt werden, die intensiv zur Analyse von Kohlenhydraten in der MS verwendet werden. Wie in der Einleitung beschrieben, kann es zur besseren Freisetzung der Glykane aus der Matrix mittels der MALDI-Technik nötig sein, das Glykane an dem reduzierenden Ende derivatisiert werden [88]. Die Eingabe der Derivatisierung soll für den Benutzer einfach möglich sein, und es sollte auch ohne Probleme möglich sein, die

Derivatisierung zu ändern. Daher ist es vorgesehen, eine Derivatisierung wie ein normales Residuum einzugeben (siehe Abbildung 37).

```
b-D-Galp-(1-4)-b-D-GlcNAc-(1-3)-b-D-Galp-(1-4)-Hex-2AP
                        |
       a-L-Fucp-(1-3)+
```

Abbildung 37: Derivatisierung von Lewis^x mit Aminopyridin

Da die MS nicht zwischen Stereoisomeren wie Galaktose und Glucose mit identischer Masse unterscheiden kann, wird oft die Komposition eines Monosaccharid in Form z.B. von Hexose (Hex) oder Hexosamin (HexNac) etc. zu seiner Bezeichnung verwendet.
Der Algorithmus zur Zerlegung der Glykane in seine Fragmente basiert, wie im vorigen Abschnitt, auf einer Bibliothek von Templaten von Monosacchariden und anderen Molekülbausteinen. Die Anzahl der benötigten Template würde sehr stark ansteigen, wenn man je einen weiteren Baustein für jede Hexose wie z.B. □-D-Manp-2AP und □-L-Glucp-2AP anlegen müsste. Deshalb haben wir uns entschieden, Derivatisierungen nur in Kombination mit der Bezeichnung für Kompositionen zu ermöglichen. Für die oben genannten Hexosen ergibt sich daher ein Template mit der Bezeichnung Hex-2AP.

4.12.3 Eingabe der Kohlenhydrate

Die digitale Eingabe von Zuckerstrukturen ist aufwendig, da durch die strukturelle Vielfalt der Kohlenhydrate die Art der Residuen sowie ihre Verknüpfungen und Verzweigungen eingegeben werden müssen. In der *CARBBANK* wurde dazu die erweiterte IUPAC-Nomenklatur für Kohlenhydrate[22, 23] verwendet. Diese Form der Eingabe wird konsequent für alle Algorithmen, Programme, Web-Tools und Datenbanken, die in der Zentralen Spektroskopie entwickelt wurden, verwendet, um dem Benutzer eine einheitliche und einfache Eingabemöglichkeit zu bieten.

```
   a-L-Fucp-(1-2)+
                 |
            b-D-Galp-(1-4)-b-D-GlcpNac-(1-3)+
                 |                          |
a-D-GalpNAc-(1-3)+                          |
                                            |
                        b-D-Galp-(1-4)-b-D-GlcpNAc-(1-3)-b-D-Galp
                                            |
   a-L-Fucp-(1-2)+                          |
                 |                          |
           b-D-Galp-(1-4)-b-D-GlcpNAc-(1-6)+
                 |
a-D-GalpNAc-(1-3)+
```

Abbildung 38: Notation für die Eingabe eines verzweigten Kohlenhydrats

Zusätzlich ist darauf zu achten, dass grundsätzlich angegeben werden muss, ob es sich bei der Verknüpfung am anomeren Kohlenstoffatom um eine □ oder □ verknüpfte glykosidische Bindung handelt. Substitutionen einzelner Hydroxylgruppen sind ebenfalls möglich. So erhält man aus der Eingabe von

a-D-GlcpNSO3

ein Glucosamin, bei dem das Amin an der Position 2 mit einer Sulfat-Gruppe substituiert ist. Durch die Verwendung von Substituenten lassen sich ohne Probleme auch recht komplexe Glykane und Heparine eingeben, die dann von den Programmen verarbeitet werden können. Einige Beispiele habe ich in der folgenden Tabelle zusammengestellt:

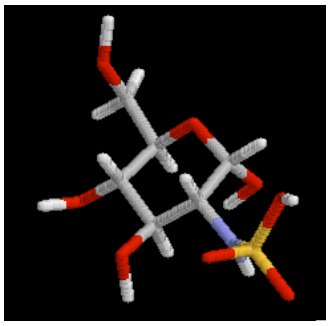
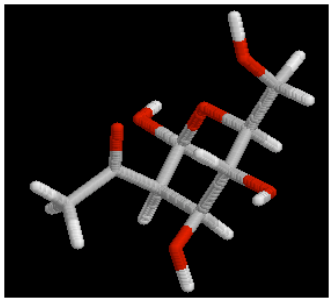
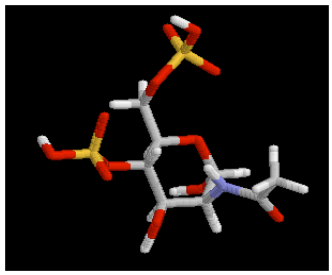
Eingabe	Struktur
a-D-GlcpNSO3	
b-L-Manp2Ac	
a-D-ManpNAc4SO36SO3	

Tabelle 10: Beispiele für die Eingabe von Substituenten

4.12.4 Eingabe von Kompositionen

Leider lassen sich mit Hilfe der MS keine sterischen Unterschiede bei den untersuchten Zuckerresiduen bestimmen. So besteht keine Möglichkeit eine Glucose von einer Galaktose zu unterscheiden, da die Massen der entstehenden Ionen in dem resultierenden Spektrum gleich sind. Daher werden vielfach nur Kompositionen ermittelt. Auch diese Möglichkeit musste bei der Eingabe Berücksichtigung finden. So wurde der Code ‚hex' für eine einzugebende Hexose festgelegt und für Residuen, die z.B. einem N-Acetyl-Galaktosamin entsprechen sollen, der Code ‚hexNac' festgelegt. Man kann also das obige Beispiel auch ohne Probleme in der folgenden Form eingeben:

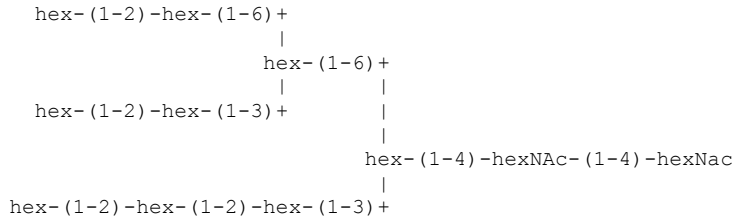


Abbildung 39: Notation für die Eingabe einer Komposition

4.12.5 Eingabe von persubstituierten Verbindungen

Kohlenhydrate sind im Gegensatz zu Proteinen nicht so leicht zu ionisieren. Daher wird oft versucht, durch eine chemische Derivatisierung die Empfindlichkeit einer Verbindung zu erhöhen.
Sehr oft werden deshalb die Hydroxylgruppen von Kohlenhydraten durch Reaktion mit Trimethyl-Silan oder anhydrierter Essigsäure permethyliert beziehungsweise peracetyliert. Dabei werden sämtliche freie Hydroxylgruppen in einen Methylether oder Essigsäurerest überführt. Die Persubstituierung wird sehr oft bei Untersuchungen verwendet, bei denen FAB als Anregungsart zur Gewinnung der Ionen verwendet wird. Bei der Bildung von Fragmenten, die aus dem Bindungsbruch von zwei Bindungen im Ring resultieren, werden charakteristische A- und X-Ionen gebildet, die Rückschlüsse über die beiden Kohlenstoff-Atome gestatten, die an der glykosidischen Bindung beteiligt sind.
Für die tägliche Arbeit mit *GLYCO-FRAGMENT* muss es möglich sein, persubstituierte Zuckerstrukturen sehr einfach einzugeben: dazu wird einfach an die jeweiligen Residuen ein ‚pMe' für eine Permethylierung beziehungsweise ‚pAc' für eine Peracetylierung angehängt. Dieses wird vom Programm *GLYCO-FRAGMENT* erkannt und automatisch werden diese Gruppen hinzugefügt. Sollte der seltene Fall eintreten, dass ein Zuckerresiduum unterschiedlich substituiert ist, so ist eine explizite Angabe der jeweiligen Substitution erforderlich. Die Eingabe

a-D-Glcp2Me3Me4Ac5Me

würde zum Beispiel eine □-D-Glucose repräsentieren, die an den Hydroxylgruppen der Kohlenstoffatome C2, C3, C5 mit einen Methylrest substituiert sind und an der Hydroxylgruppe des C4 eine Acetylierung erfolgt ist.

4.12.6 Eingabe von chemischen Derivaten

Da Kohlenhydraten ein Chromophor fehlt, sind sie in der Regel sehr schwer mittels der gebräuchlichen UV/VIS-Detektoren bei der HPLC zu detektieren. Daher werden Glykane häufig am reduzierenden Ende derivatisiert. Die gebräuchlichsten Substanzen, die zur Derivatisierung[89] eingesetzt werden, sind in Tabelle 11 aufgelistet.

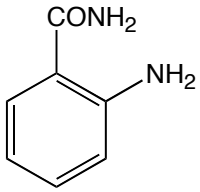
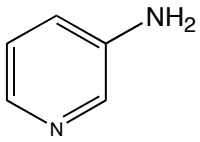
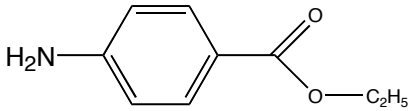
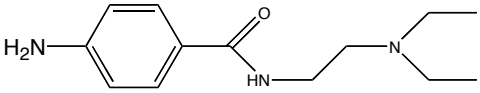
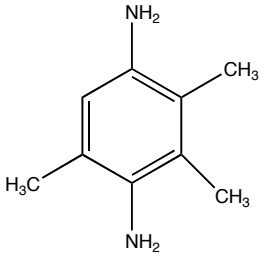
Derivativ	Abkürzung	Struktur
2-Amino-Benzamid	2-AB	
2-Aminopyridin	2-AP	
4-Aminobenzoesäurethylester	4ABEE	
4-Amino-N-(2-diethylaminoethyl)-Benzamid	DEAEAB	
Trimethyl-4-Amino-Anilin	4TMAPA	

Tabelle 11: gebräuchliche Substanzen zur Derivatisierung

Die MS-Spektren zeigen durch die Derivatisierungen ebenfalls eine erhöhte Tendenz Ionen zu bilden. Von verschiedenen Arbeitsgruppen wurde untersucht, wie die Nachweisbarkeit von Glykanen mittels MS durch Derivatisierungen verbessert werden kann. So zeigen Untersuchungen, dass durch die Derivatisierungen von Glykanen mit 2-amino-Pyridin[88, 90], 2-(diethylamino)ethyl-4-amino-Benzoesäureethylester[42], 4-Aminobenzoesäurethylester[88], 2-Aminobenzamid[53] und Trimethyl-4-Amino-Anilin[91], Massenspektren resultieren, die wesentlich besser interpretierbare Massen aufweisen als dies bei nicht derivatisierten Zuckern der Fall ist.

Natürlich müssen auch diese Derivatisierungen von dem *GLYCO-FRAGMENT*-Algorithmus bearbeitet werden können. Wichtig war jedoch, dass die systematische Bezeichnung und Linkage-Information für die einzelnen Ionen weiterhin der Nomenklatur von Domon und Costello folgen, und die Derivate nicht vom Algorithmus als weitere Residuen betrachtet werden. Zur Eingabe von derivatisierten Zuckern wird anstelle des letzten Residuums einfach die Komposition plus Derivatisierung eingegeben. Hex-2AP bezeichnet eine Derivatisierung des Glykans mit 2-Aminopyridyn.

```
b-D-Galp-(1-4)-b-D-GlcNAc-(1-3)-b-D-Galp-(1-4)-Hex-2AP
                           |
        a-L-Fucp-(1-3)+
```

Abbildung 40: Notation für die Eingabe einer Derivatisierung

Mit Hilfe dieser Eingabeoptionen ist eine sehr gute Möglichkeit geschaffen, die häufigsten Fragestellungen in der massenspektrometrischen Analytik effizient auswerten zu können.
Der Algorithmus steht allen interessierten Wissenschaftlern über ein webbasiertes Interface zur Verfügung und kann unter der URL: http://www.dkfz-heidelberg.de/spec/projekte/fragments/ angesprochen werden. Dort sind auch weitere Beispiele für all die hier besprochenen Eingabemöglichkeiten verfügbar.

4.12.7 Technische Umsetzung

Eine der wesentlichen Anforderungen an das Interface für den *GLYCO-FRAGMENT* Algorithmus bestand darin, dass die komplette Benutzerinteraktion über Web-Browser erfolgen soll. Als Webserver kam der am DKFZ zurzeit verwendete Apache-Webserver[1] in der Version 1.3.24 zum Einsatz. Dieser Webserver läuft sehr stabil und kann auf der Serverseite sehr leicht durch Module, wie das hier verwendete PHP erweitert werden. PHP[2] wurde in der aktuellen Version 4.3 verwendet. Die Berechnung der Fragment-Ionen, als der die Geschwindigkeit bestimmende Schritt, sollte quasi interaktiv und im Bereich von Millisekunden erfolgen, daher war eine reine Lösung auf der Basis des PHP-Interpreters wenig sinnvoll. Es wurde daher ein CGI-Skript entwickelt, das die Übergabe der Daten an das eigentliche Programm realisiert, das in der Programmiersprache C entwickelt wurde. Dadurch erfolgt die Berechnung der einzelnen Fragmente in Sekundenbruchteilen und die berechneten Fragmente können quasi in Echtzeit ausgegeben werden.

4.12.8 Das Webinterface

Zur Interaktion mit dem Benutzer musste ein Webinterface entwickelt werden, das es gestattet, auf einfache Weise alle benötigten Informationen an das eigentliche Programm zu übermitteln.

[1] http://www.apache.org/
[2] http://www.php.net/

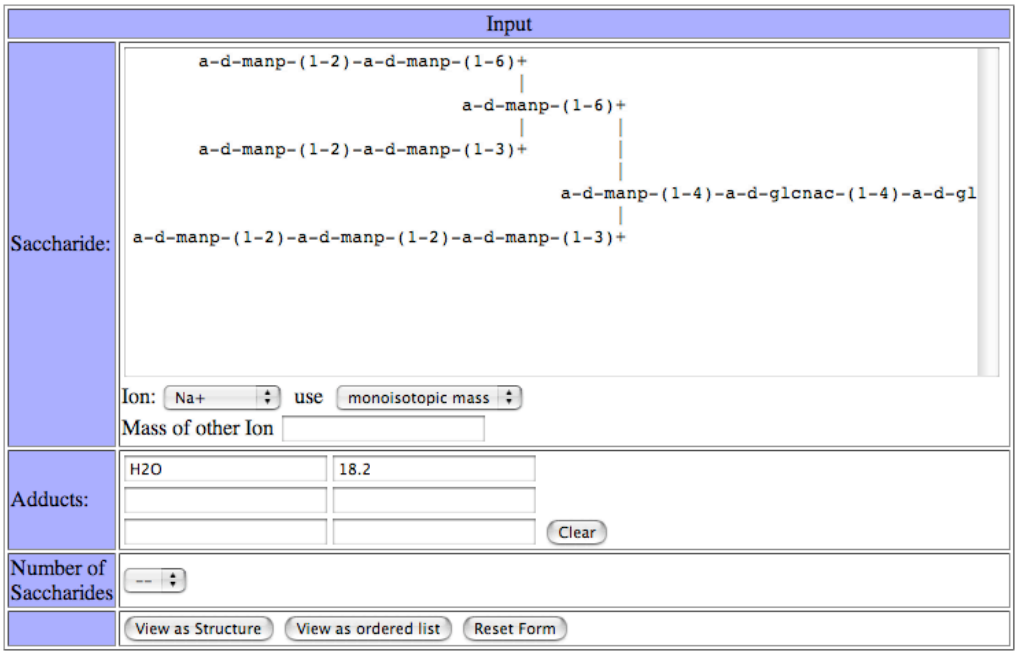


Abbildung 41: Das Webinterface des Programms *GLYCO-FRAGMENT*

Das Webinterface unterteilt sich in vier verschiedene inhaltliche Bereiche. Im oberen Fenster erfolgt die Eingabe der Struktur. Es handelt sich hierbei um einen einfachen ASCII-Editor. Die Zuckersequenzen können vom Benutzer über die Tastatur eingegeben werden. Es besteht auch die Option, durch einen Kopieren/Einfügen-Mechanismus die Beschreibungen aus anderen Dateien direkt einzufügen. Alle oben beschriebenen Formate werden akzeptiert. Die Verwendung des ASCII-Editors hat sich auch bei der Analyse von Serien von Strukturen bewährt.
Die formale Korrektheit der eingegebenen Struktur wird bei der Eingabe auf zwei verschiedene Arten kontrolliert. Zum einen wird durch das Programm *SWEET-II* die Eingabe auf Syntaxfehler geprüft. Auch werden nur solche Residuen akzeptiert, die in der *SWEET-II* Datenbasis enthalten sind. Die bei der Struktureingabe gefundenen Probleme werden dem Benutzer präsentiert und er kann an Hand der Fehlermeldung die eingegebene Struktur überprüfen. Zur Sicherheit wird der Benutzer gezwungen, die Anzahl der in der eingegebenen Struktur enthaltenen Residuen anzugeben.

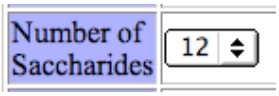


Abbildung 42: Pulldown-Menü zur Eingabe der Residuenzahl

Es erfolgt dann eine Kontrolle mit der von *GLYCO-FRAGMENT* gefundenen Anzahl an Residuen. Sollte eine Abweichung festgestellt werden, wird der Benutzer mit der folgenden Fehlermeldung

```
Wrong number of residues found! I have found 11 residue(s).
Should have found 12 Residue(s).
```

auf inkonsistente Daten hingewiesen. Auch hier muss der Benutzer seine Struktureingabe überprüfen und die entsprechenden Änderungen vornehmen.
Unterhalb dieses Eingabefeldes können die Angaben zur verwendeten Messmethode gemacht werden. Für das Programm wichtig sind die zur Ionisierung verwendete Atomart und die gemessene Signalart.

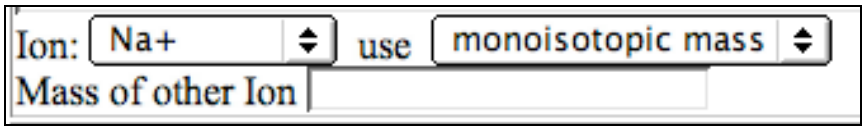


Abbildung 43: Eingabefelder für die Signalart und das verwendete Ion

Durch die beiden Pulldown-Menüs kann sehr schnell das verwendete Ion ausgesucht werden und aus den Signalarten monoisotopische und durchschnittliche Masse gewählt werden. Sollte ein Ion verwendet werden, das sich nicht in der Liste befindet, kann in das Eingabefeld <Mass of other Ion> der Massenwert direkt eingetragen werden.
Da es bei bestimmten Substanzklassen immer wieder zur Abspaltung von bestimmtem Atomgruppen wie eines H_2O oder eines Acetyl-Restes kommen kann, besteht die Möglichkeit bis zu drei verschiedene Addukte anzugeben, die dann von allen berechneten Massen subtrahiert werden.

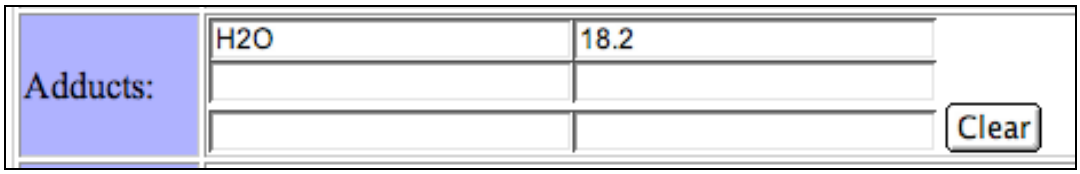


Abbildung 44: Eingabefelder für bis zu drei verschiedene Addukte

4.12.9 Ausgaben des Webinterfaces

Die Ausgabe des Webinterfaces unterscheidet vier verschiedene Modi. Bei der ersten Ausgabeoption kann man sich das Ergebnis des Programms *GLYCO-FRAGMENT* in einer Form ansehen, die die eingegebene Struktur des Glykans beibehält. Die anderen drei Modi präsentieren die berechneten Massen in einer Liste und ermöglichen so eine einfache Zuordnung seiner gemessenen Peaks. Die gewünschte Ausgabe kann mit Hilfe der vier Buttons im unteren Bereich des Webinterface ausgewählt werden.

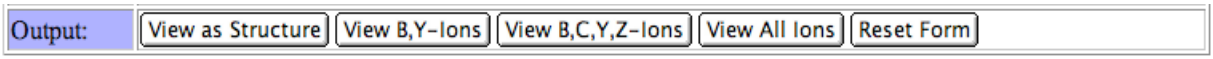


Abbildung 45: Button-Leiste des Webinterfaces

Mit dem ersten Button ist es möglich, sich eine Ausgabe der Ergebnisse in einer Form anzusehen, die sich an der eingegebenen Struktur orientiert. Die anderen drei Buttons geben die Ergebnisse in einer mehr oder weniger langen Liste an.

4.12.10 Ergebnisdarstellung als Struktur

Zur Darstellung der Ergebnisse in der Strukturansicht, wird die eingegebene Struktur analysiert und es werden in den ausgegebenen HTML-Code Tags eingebaut, die es ermöglichen durch einfaches Überfahren mit der Maus weitere Informationen zu erhalten.

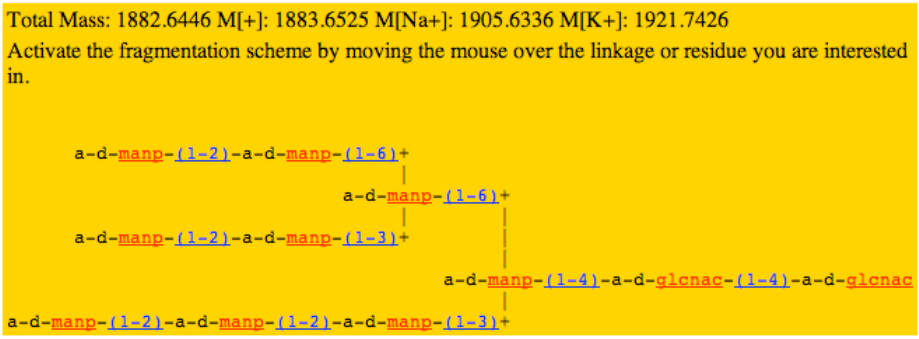


Abbildung 46: Ergebnisdarstellung als Struktur

Durch Bewegen der Maus über einzelne Residuen und glykosidische Bindungen können nun über eingeblendete Layer die eigentlichen Werte für die Ionen abgefragt werden. Bewegt man die Maus über eine glykosidische Bindung, so erhält man die Massen der B-, C-, Y- und Z-Ionen sowie die der jeweiligen subtrahierten Addukte.

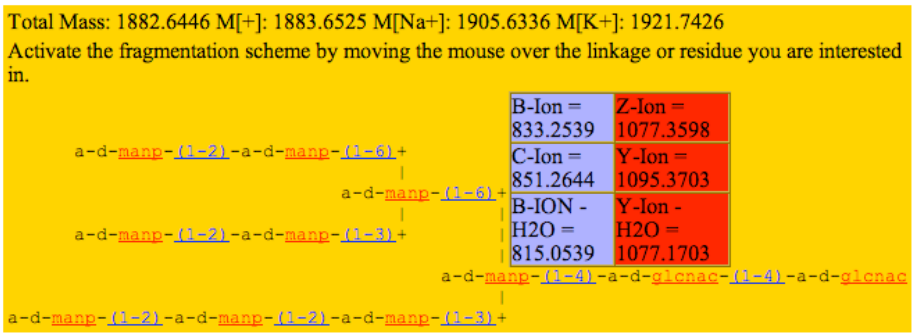


Abbildung 47: Layer mit B-, C-, Y- und Z-Ionen

Analog dazu erhält man die Liste aller möglichen A- und X-Ionen, wenn man die Maus über ein Residuum in der Darstellung bewegt. Diese Darstellung ist sehr wichtig, da gerade so die an der glykosidischen Bindung beteiligten Atome sehr schnell herausgefunden werden können.

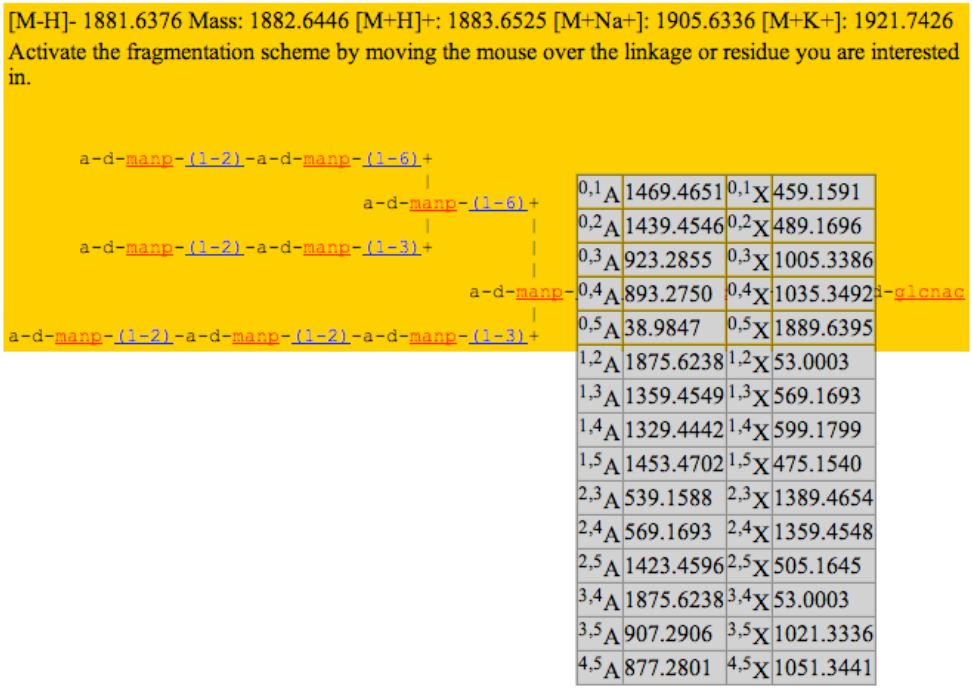


Abbildung 48: Layer mit den zugehörigen A- und X-Ionen

Mit dieser Darstellung hat man die Möglichkeit gezielt einzelne Massenwerte zu ermitteln, die es einem gestatten, Strukturinformationen aus seinem MS-Spektrum zu ermitteln, da man einen direkten Überblick über die A- und X-Ionen erhält.

4.12.11 Ergebnisdarstellung als Liste

Als Alternative zu dieser strukturierten Darstellung kann man sich die aus der Struktur berechneten Ionen auch als Liste ausgeben lassen. Die Liste ist nach aufsteigenden Massenwerten sortiert und ermöglicht so eine einfache Überprüfung, ob diese Peaks auch im gemessenen Spektrum vorhanden sind. Es kann zwischen drei verschiedenen Listen gewählt werden. Für einen ersten Überblick ist es möglich, sich nur die B- und Y-Ionen anzusehen. Die Liste besteht aus zwei Spalten. In der ersten befindet sich die berechnete Masse, und in der zweiten Spalte befinden sich die Informationen der Ionen, die diesem Massenwert entsprechen können. Die Benennung der Ionen erfolgt nach der in der Einleitung beschriebenen Nomenklatur. Zur Steigerung der Übersichtlichkeit werden alle Ionen mit demselben Massenwert in einer Zeile zusammengefasst.

[M-H]- 1881.6376 Mass: 1882.6446 [M+H]+: 1883.6525 [M+Na+]: 1905.6336 [M+K+]: 1921.7426	
Mass in amu	Ion
185.0426	$B_{4,4,3,2,2}$ + Na+
	$B_{4,4,6,3,2}$ + Na+
	$B_{4,4,6,6,2}$ + Na+
244.0797	Y_4 + Na+
347.0954	$B_{4,4,3,2}$ + Na+
	$B_{4,4,6,3}$ + Na+
	$B_{4,4,6,6}$ + Na+
447.1591	$Y_{4,4}$ + Na+
509.1482	$B_{4,4,3}$ + Na+
833.2539	$B_{4,4,6}$ + Na+
1095.3703	$Y_{4,4,6}$ + Na+
1419.4760	$Y_{4,4,3}$ + Na+
1481.4651	$B_{4,4}$ + Na+
1581.5288	$Y_{4,4,3,2}$ + Na+
	$Y_{4,4,6,3}$ + Na+
	$Y_{4,4,6,6}$ + Na+
1684.5445	B_4 + Na+
1743.5816	$Y_{4,4,3,2,2}$ + Na+
	$Y_{4,4,6,3,2}$ + Na+
	$Y_{4,4,6,6,2}$ + Na+

Abbildung 49: Ergebnisliste mit B- und Y-Ionen

Will man mehr Informationen erhalten, so kann man die Liste auch mit den C- und Z-Ionen erweitern. Das Entstehen dieser Ionen hängt aber stark von den Messbedingungen ab. In der dritten Form der Listendarstellung können auch die A- und X-Ionen mit angegeben werden. Diese Liste ist allerdings sehr lang und sollte erst ausgewählt werden, wenn man sich um die Aufklärung, der an den glykosidischen Bindungen beteiligten Kohlenstoffatome, bemühen will.

4.12.12 Beispiele

Mit diesem Web-Tool ist eine Analyse von sehr unterschiedlichen Kohlenhydratgruppen möglich. So können zum einen alle O-, und N-Glykane in ihre Fragmente zerlegt werden, aber auch die Klasse der Glykolipide und Lipopolysaccharide kann analysiert werden. Dieses soll in den folgenden Beispielen gezeigt werden. Dazu werden auch die unterschiedlichen Möglichkeiten der Darstellung gezeigt und auch die Möglichkeit der Verwendung von Derivaten.

4.12.12.1 Fragmentierung eines N-Glykans

Als erstes Beispiel zeige ich hier die Fragmentierung eines N-Glykans. N-Glykane sind eine sehr gut untersuchte Klasse der Protein-Glykosylierungen[25, 36, 92], zu der es eine große Anzahl von Publikationen im Bereich der massenspektrometrischen Untersuchungen gibt. Leider ist es sehr schwierig, reale Massenspektren zu erhalten. Zur Überprüfung des obigen Algorithmus wurden daher Spektren verwendet, die sehr gut aus der Literatur übernommen werden konnten [54, 93].
In diesen Arbeiten wurde unter anderem ein triantennäres Glykan vom High-Mannose-Typ verwendet, dessen Sequenz in der Abbildung 50 zu sehen ist.

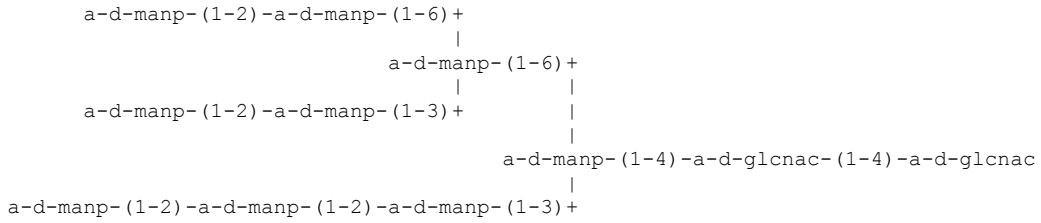


Abbildung 50: Struktur des gesuchten Beispielglykans

Dieses Glykan wurde nun kristallisiert und mittels eines mit 10 Hz gepulsten Lasers verdampft und mit einem TOF-Massenanalysierer die Spektren aufgenommen. Man erhielt das folgende Spektrum:

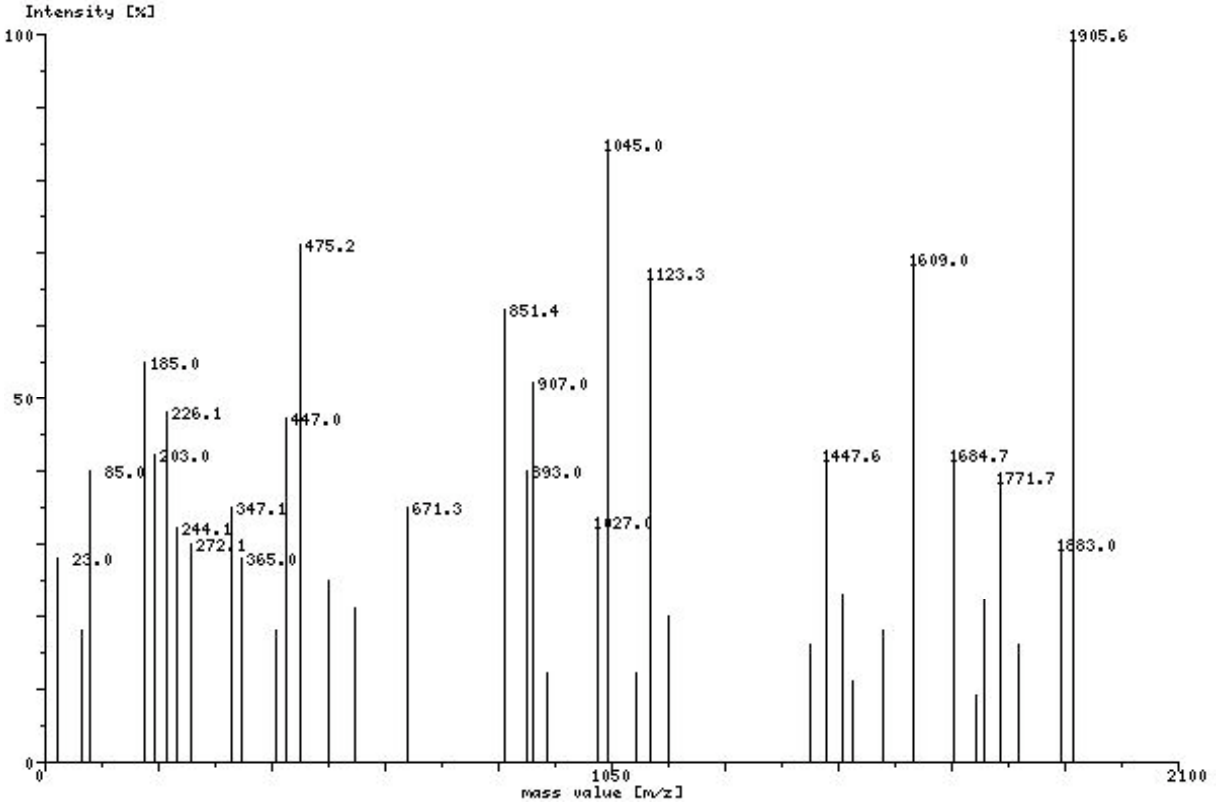


Abbildung 51: Spektrum des High-Mannose N-Glykans

4.12.12.2 Analyse des Spektrums mit dem Programm *GLYCO-FRAGMENT*

Es sollte nun versucht, mit dem Programm *GLYCO-FRAGMENT* eine Kohlenhydrat-Sequenz zu finden, die eine sehr gute Übereinstimmung mit diesem Spektrum zeigte. Durch Ausprobieren verschiedener Strukturen konnte nach längerer Zeit die obige Struktur zugeordnet werden. Die in dem Massenspektrum enthaltenen Peaks konnten wie folgt zugeordnet werden:

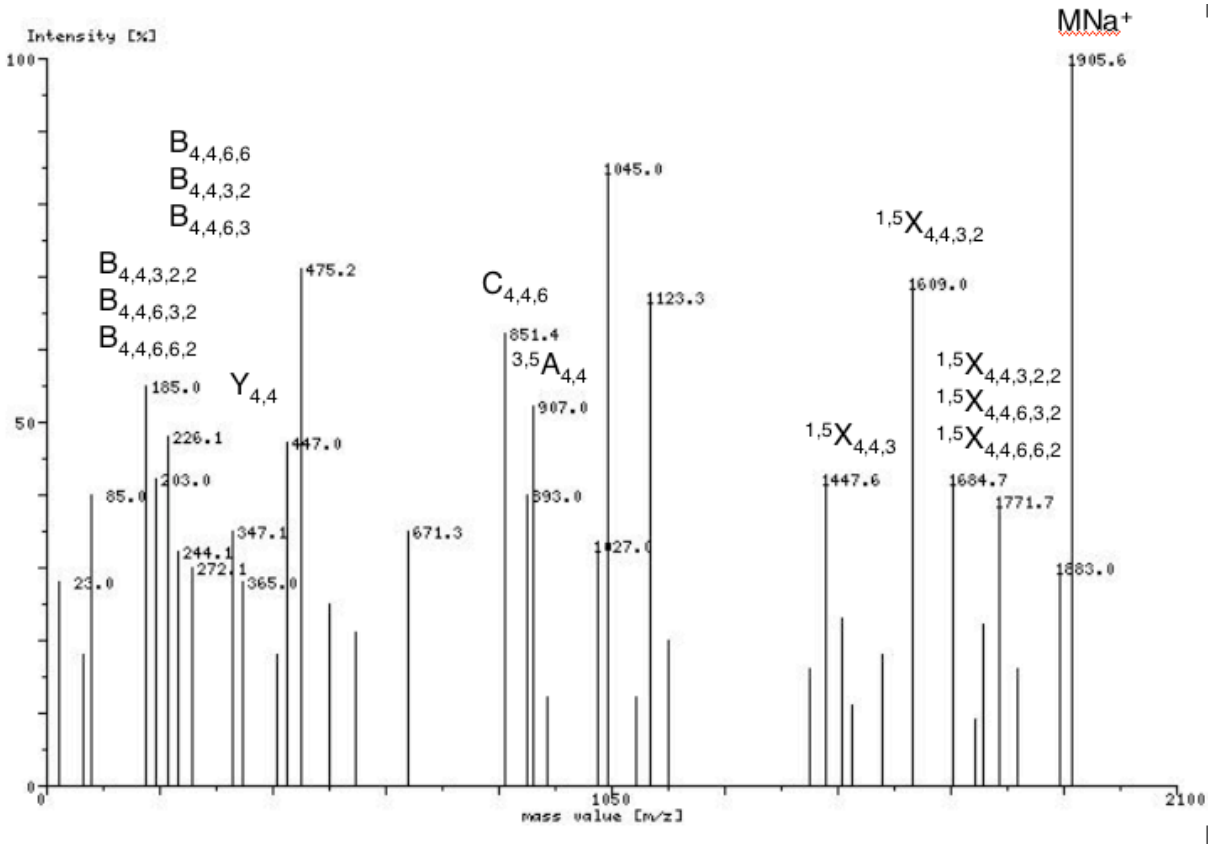


Abbildung 52: Spektrum aus Abbildung 51 mit den zugeordneten Peaks

Es ist sehr schön zu sehen, dass gerade die ^{1,5}X-Ionen sehr zahlreich vorhanden sind und Aussagen über die einzelnen Atome gestatten, die an den glykosidischen Bindungen beteiligt sind. ^{1,5}X-Ionen entstehen vorzugsweise bei 1-2 verknüpften Zuckerresiduen. Gerade bei Zuckerresiduen, von denen mehrere Seitenketten verzweigen, können so Aussagen über die beteiligten Kohlenstoffatome und die Zusammensetzung der Ketten gemacht werden. Natürlich können nicht alle im Spektrum enthaltenen Peaks zugeordnet werden, da viele Fragmente sich nicht nach den Regeln von Domon und Costello bilden.

4.12.12.3 Fragmentierung eines derivatisierten Oligosaccharids

Als weiteres Beispiel zeige ich hier die Fragmentierung eines derivatisierten Oligosaccharids[88]. In dieser Arbeit wurde unter anderem das folgende Oligosaccharid, dessen Sequenz in der Abbildung 53 zu sehen ist, untersucht:

```
a-d-manp-(1-4)-a-d-manp-(1-4)-a-d-manp-(1-4)-a-d-manp-(1-4)-hex-2Ap
```

Abbildung 53: Struktur des derivatisierten Oligosaccharids

Dieses Oligosaccharid wurde in einer 2,5-Dihydroxybenzoesäure Matrix kristallisiert und mittels eines MALDI/TOF Massenspektrometers die Spektren aufgenommen. Man erhielt das folgende Spektrum:

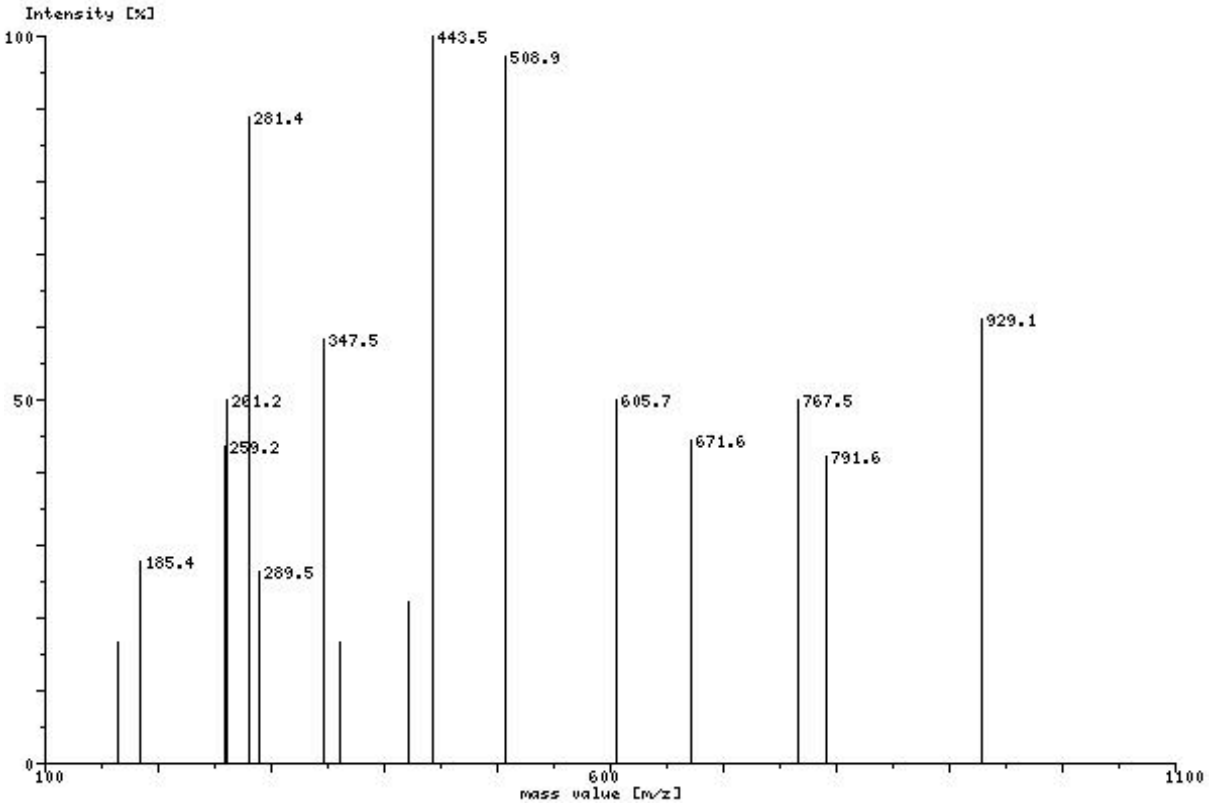


Abbildung 54: Spektrum des derivatisierten Oligosaccharids

4.12.12.4 Analyse des Spektrums mit dem Programm *GLYCO-FRAGMENT*

Es sollte nun auch mit diesem Spektrum versucht werden mit dem Programm *GLYCO-FRAGMENT* eine Kohlenhydrat-Sequenz zu finden, die eine sehr gute Übereinstimmung mit diesem Spektrum zeigte. Durch Ausprobieren verschiedener Strukturen konnte die obige Struktur zugeordnet werden. Die in dem MS-Spektrum enthaltenen Peaks konnten wie folgt zugeordnet werden:

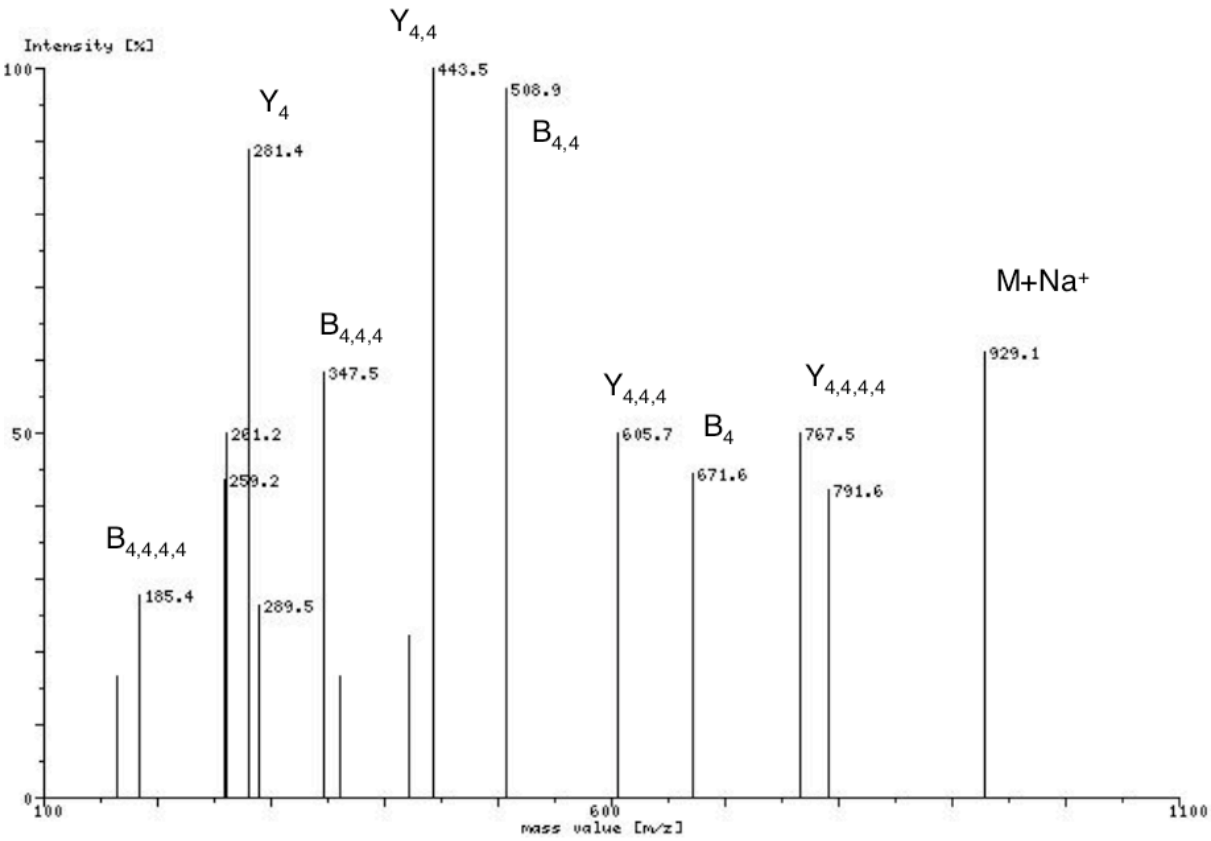


Abbildung 55: Spektrum aus Abbildung 54 mit den zugeordneten Peaks

Es ist sehr schön zu sehen, dass ohne Probleme alle B- und Y-Ionen zugeordnet werden konnten.

4.12.12.5 Fragmentierung eines Gangliosids

Als letztes Beispiel zeige ich hier die Fragmentierung eines Gangliosids. Die Ganglioside wurden von Prof. Peter-Katalinic in Münster untersucht[77]. In diesen Arbeiten wurde unter anderem das Gangliosid untersucht, dessen Sequenz in der Abbildung 56 zu sehen ist.

```
NeuNAc-(1-3)-a-d-Galp-(1-4)-a-d-GlcpNac-(1-4)-a-d-Galp-(1-4)-a-d-galp-(1-1)-Ceramide
                                        |
                          a-L-Fucp-(1-3)+
```

Abbildung 56: Struktur des gesuchten Gangliosids

Das Gangliosid wurde nun mittels eines Esi-Massenspektrometers untersucht. Das Spektrum wurde im negativen Modus aufgenommen. Dieses bedeutet, dass in dem Spektrum nur Ionen mit einer negativen Ladung enthalten sind.

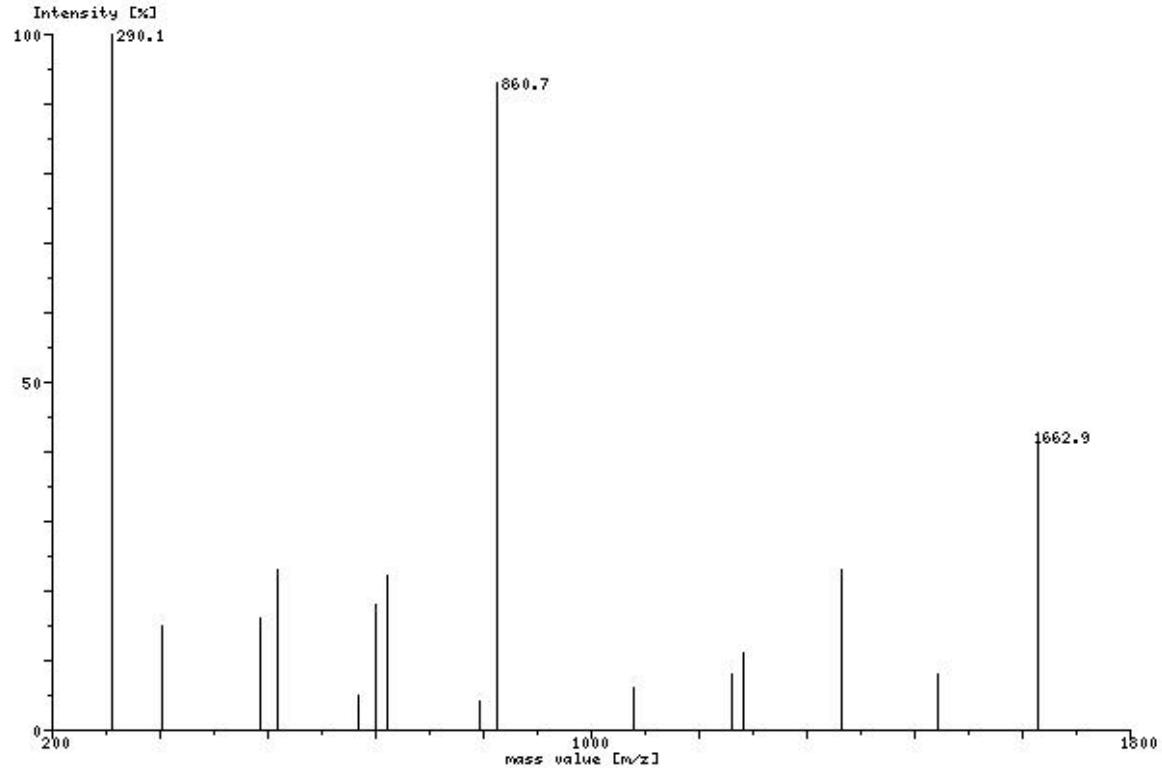


Abbildung 57: Spektrum des Gangliosid

4.12.12.6 Analyse des Spektrums mit dem Programm *GLYCO-FRAGMENT*

Durch Ausprobieren verschiedener Strukturen konnten viele Massen den Fragment-Ionen der obigen Struktur zugeordnet werden:

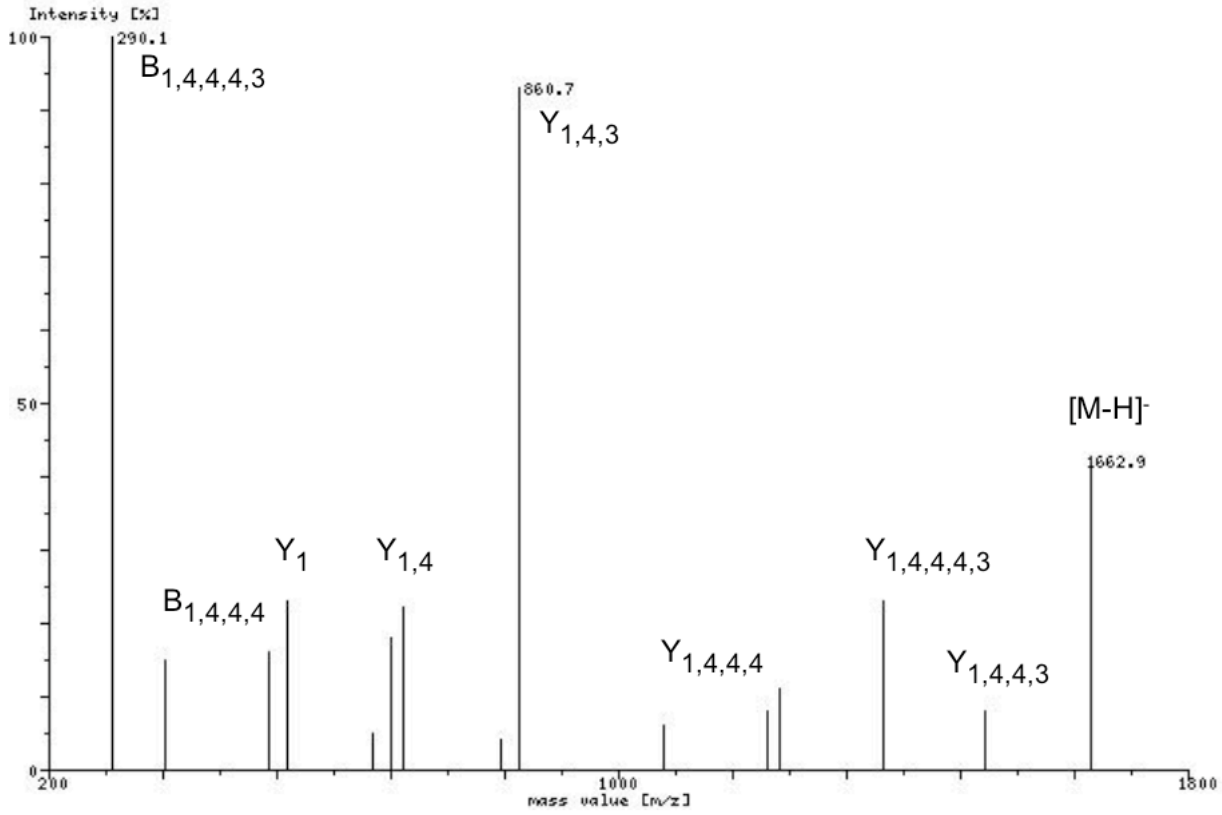


Abbildung 58: Spektrum aus Abbildung 57 mit den zugeordneten Peaks

Leider waren in dem Spektrum nicht alle B-Ionen enthalten, es konnten aber alle Y-Ionen gefunden werden.

Substanz-Klasse	Anzahl
Oligosaccharid	12
N-Glykan	19
O-Glykan	17
Lipopolysaccharid	3
Derivatisiertes N-Glykan	2
Heparinoide	2

Tabelle 12: Auflistung der untersuchten Saccharid-Strukturen

Insgesamt wurden mit dem Algorithmus 53 Strukturen aus der Literatur untersucht und konnten auch zugeordnet werden. Neun dieser Zuordnungen sind unter der oben genannten URL zu finden.

4.13 Peak-Assignment

Bei der Auswertung der gemessenen Spektren ist es sehr wichtig, auch eine vollständige Zuordnung der Peaks zu ermöglichen. So ist es sehr unbefriedigend, Peaks, die nur eine geringe Intensität haben, aber für die Aufklärung der Struktur sehr wichtig sind, unberücksichtigt zu lassen. So sind es gerade die A- und X-Bruchstücke, die eine hohe Aussagekraft für die Art der glykosidischen Bindung besitzen. Eine manuelle Zuordnung dieser Peaks ist mehr als unbefriedigend. Bei dem vorhergehenden Programm *GLYCO-FRAGMENT* haben die so genannten internen Fragmente noch keine wesentliche Rolle gespielt, da sie für die Sequenzermittlung keine hohe Aussagekraft im Vergleich zu den A-, B-, C-, X-, Y- und Z-Ionen besitzen. Bei den internen Fragmenten handelt es sich um Fragmente, die durch Spalten zweier glykosidischer Bindungen entstehen. Dieses habe ich in der folgenden Struktur verdeutlicht:

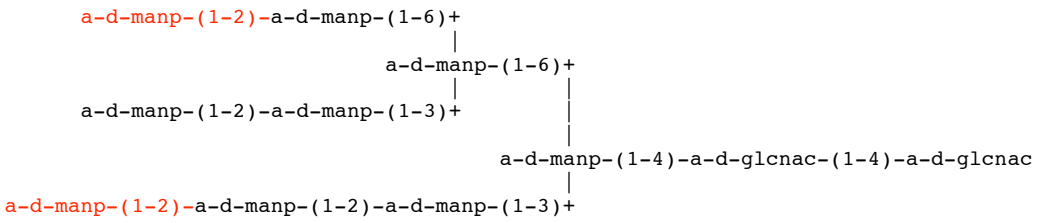


Abbildung 59: Struktur des gesuchten Beispielglykans

Durch das Abspalten zweier endständiger □-D-Mannosen entsteht nun das folgende interne Fragment:

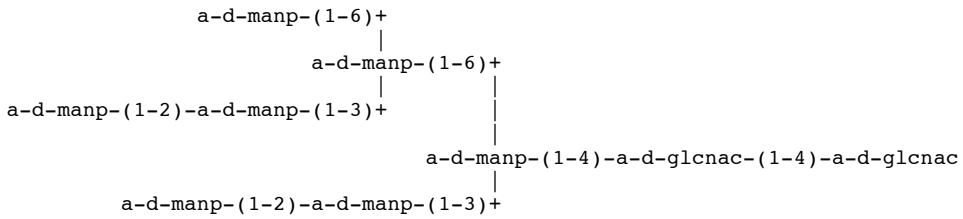


Abbildung 60: Struktur des gesuchten Beispielglykans

Dieses interne Fragment erhält nun bei der Zuordnung durch das Programm *PEAKASSIGN* die folgende Bezeichnung

$_{4,4,6,6,2}$Innerfragment$_{4,4,3,2,2}$

Die tiefgestellten Indizes entsprechen hier wieder der Linkage-Information zu den Residuen, die abgespalten worden sind.
Diese internen Fragmente gehören zu einer vollständigen Auswertung eines Massenspektrums. Dieses führte zu der Entwicklung des Programms *PEAKASSIGN*, das es ermöglichen soll, einem eingegebenen Spektrum alle möglichen internen- und Fragment-Ionen zuzuordnen. Die Berechnung der Fragment-Ionen erfolgt genauso wie oben beschrieben beim Programm *GLYCO-FRAGMENT*. Zusätzlich werden aber noch die internen Fragment-Ionen berechnet, die durch das Spalten zweier glykosidischer Bindungen entstehen. Auch dieses Programm wird webbasiert angeboten und ist unter der folgenden URL erreichbar: http://www.dkfz-heidelberg.de/spec/projekte/fragments/assign.php4.

4.13.1 Eingabe

Die Eingabe erfolgt ähnlich wie bei dem Tool *GLYCO-FRAGMENT*. Als erstes und wichtigstes muss die untersuchte Struktur eingegeben werden. So müssen auch hier das benutzte Esi-Ion und die Art der bestimmten Massen angegeben werden. Im nächsten Eingabe-Block werden die Spektren-Datei und der zugehörige erlaubte Fehler eingegeben.

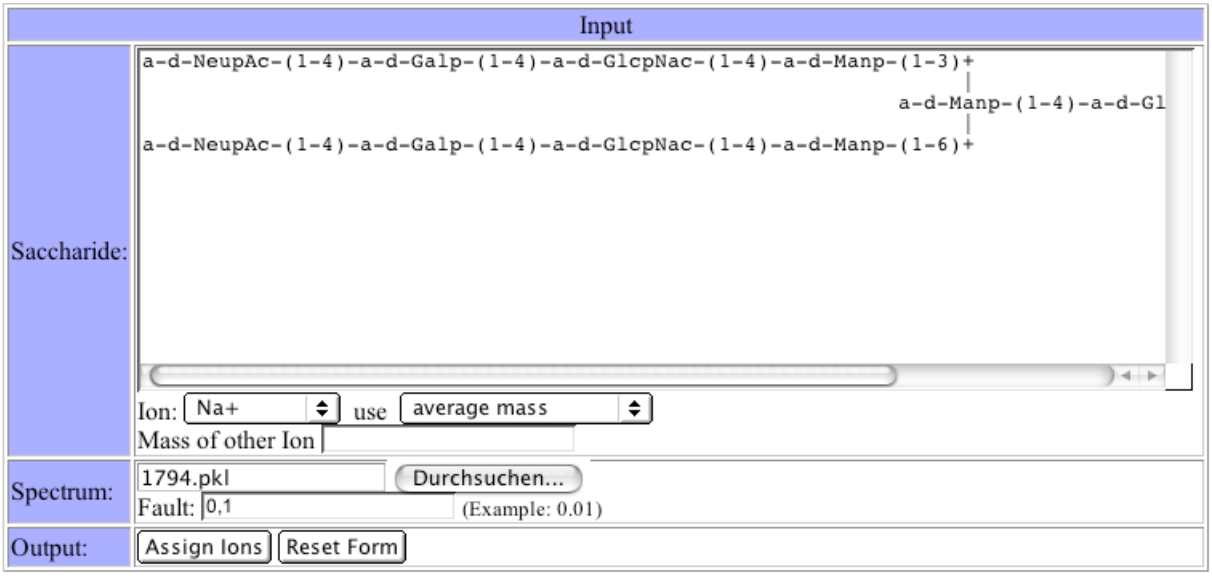


Abbildung 61: Eingabe der benötigten Parameter

Im Moment werden von diesem Tool zum einen das Micromass pkl-Format unterstützt, zum andern aber auch das Sequest dta-Format. Es wird so eine große Menge an im Moment üblichen Formaten unterstützt. Andere Formate werden zurzeit nicht unterstützt. Die meisten Programme, die vom Hersteller der Massenspektrometer mitgeliefert werden, bieten aber eine Export-Funktion in eines dieser Formate. Bei dem hier verwendeten Beispiel handelt es sich um ein biantennäres N-Glykan[82], dessen MS-Spektrum freundlicherweise von Hélène Perreault, University of Manitoba zur Verfügung gestellt worden ist. Es handelt sich um ein ESI-Spektrum, bei dem die durchschnittlichen Massen mit Na^+ als Esi-Ion gemessen wurden.

4.13.2 Ausgabe

Nach Absenden dieser Informationen erhält man alle Peaks des Spektrums mit den zugehörigen Intensitäten in Form einer Liste angezeigt. In einer dritten Spalte sind die von dem Programm zugeordneten Bezeichnungen angegeben.

Peak	Intens	theoretical mass/ assignment
512.1900	6.0000	
513.1960	6.0000	
524.2080	6.0000	
.		
.		
550.2110	14.0000	550.1748 $_{4,4,3,4,4,4}$Innerfragment$_{4,4,3}$+ Na+ 550.1748 $_{4,4,6,4,4,4}$Innerfragment$_{4,4,3}$+ Na+ 550.1748 $_{4,4,6,4,4,4}$Innerfragment$_{4,4,6}$+ Na+ 550.1748 ^{3,5} A$_{4,4,3,4}$ + Na+ 550.1748 ^{3,5} A$_{4,4,6,4}$ + Na+
550.2510	16.0000	
.		
.		
1427.5070	8.0000	1427.4923 Y$_{4,4,3}$ + Na+ 1427.4923 Y$_{4,4,6}$ + Na+
1427.5640	8.0000	
.		
.		
1798.7100	6.0000	
1798.8020	10.0000	

Abbildung 62: Ausgabe des *PEAKASSIGN* Tools

Dabei werden sowohl die internen Fragmente, als auch die anderen Fragmente den passenden Peaks zugeordnet. Es ist also ohne Problem möglich, auch kleinere Peaks zu finden und zuzuordnen.

4.13.3 Ergebnis und Diskussion

Mit Hilfe dieser Tools konnte gezeigt werden, wie einfach und schnell die Zuordnung einzelner Peaks in einem Massenspektrum geschehen kann. Anders als bei der manuellen Zuordnung ist es mit den drei Programmen *PEAKASSIGN*, *GLYCO-FRAGMENT* und *FINDYSERIES* ohne Probleme möglich, eine große Anzahl im Spektrum enthaltener Peaks, die nach der Nomenklatur von Domon und Costello[58] bzw. nach den Regeln der Benennung von Peptid-Fragmenten erzeugt werden können, zuzuordnen. Als erstes kann mit dem Programm *FINDYSERIES* die Sequenz eines Peptids ermittelt und untersucht werden, ob sich post- oder cotranslationale Modifikationen an dieser Peptidsequenz befinden. Sollte es sich um Glykosylierungen handeln, kann mit Hilfe des Programms *GLYCO-FRAGMENT* versucht werden, die Komposition und Sequenz dieser Modifikation zu ermitteln. Als letzter Schritt kann mit dem Programm *PEAKASSIGN* die Aussage des Programms *GLYCO-FRAGMENT* erhärtet werden, da hier auch zusätzlich interne Fragmente zugeordnet werden. Die Zuordnung erfolgt sehr schnell und die Änderung der zu ermittelnden Ausgangsstruktur kann schnell durchgeführt werden, da eine sofortige Neuberechnung des theoretischen Spektrums erfolgen kann. Bei allen Programmen hat es sich als großer Vorteil erwiesen, dass zuerst die neutralen Fragmente berechnet werden. Ausgehend von diesen Fragmenten können dann sehr schnell die Fragment-Ionen berechnet werden. Aus diesen Fragmenten können in Zukunft auch

Ionen mit einer Ladungszahl höher als eins berechnet werden. Diese Ionen finden bei den Programmen bis jetzt noch keine Berücksichtigung, da die höchste Aussagekraft von einfach geladenen Ionen ausgeht[57]. Bei allen Programmen konnte durch eine größere Anzahl von Spektren die Richtigkeit der zu Grunde liegenden Algorithmen gezeigt werden.

4.14 Automatische Auswertung von Massenspektren

Nach der Entwicklung der Programme *GLYCO-FRAGMENT* und *PEAKASSIGN* bot sich der dort implementierte Algorithmus an, aus allen in der *SWEET-DB* gespeicherten Strukturen eine Bibliothek mit theoretischen Vergleichsspektren zu berechnen. Im Idealfall kann nun mit dem gemessenen Spektrum in der Datenbank gesucht werden, und es erfolgt eine Identifizierung der gesuchten Verbindung.
Die Datenbank soll, ähnlich wie das Programm *MASCOT* für die Proteomik, in der Lage sein, eine Hilfe bei der Aufklärung von gemessenen Spektren zu sein, die mit den im Moment gebräuchlichen Methoden der Massenspektrometrie aufgenommen worden sind. Das Webinterface sollte als Erweiterung der bestehenden *SWEET-DB* implementiert werden und in der Lage sein, die Eingabe einer Peakliste zu ermöglichen. Diese Peakliste wird dann mit allen theoretischen Spektren verglichen und es soll ein Score ermittelt werden, der dem Benutzer angibt, wie gut die eingegebene Liste mit dem gefundenen Spektrum übereinstimmt.

4.14.1 Berechnung der Spektrenbibliothek

Zur Berechnung der Spektrenbibliothek wurde für alle Strukturen, die in der *SWEET-DB* gespeichert sind, mit dem Programm *GLYCO-FRAGMENT* ein theoretisches Spektrum berechnet. Diese Spektren enthalten alle A-, B-, C-, X-, Y- und Z-Fragmente, die sich nach den Regeln von Domon und Costello bilden würden.

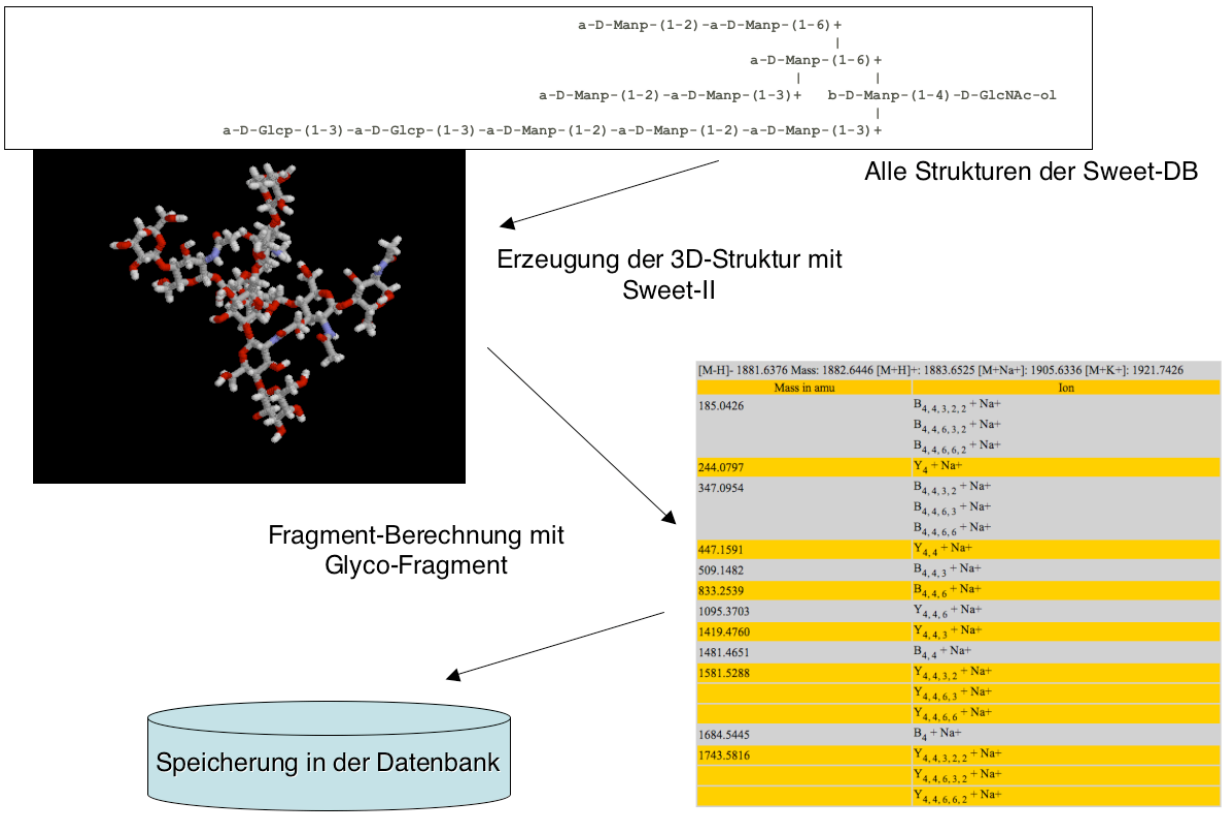


Abbildung 63: Berechnung der Spektrendatenbank

Die Berechnung der A- und X-Fragmente war besonders wichtig, da diese für die Unterscheidung von zwei Strukturen besonders aussagekräftig sind, da hier die unterschiedlichen glykosidischen Bindungen zu unterschiedlichen Fragmenten führen. Der Suchalgorithmus gestaltete sich sehr einfach. Es wird jedes Spektrum der Bibliothek mit dem eingegebenen verglichen und die besten Treffer werden ausgegeben.

4.14.2 Die Bewertungsfunktion des Suchalgorithmus

Der Algorithmus vergleicht das eingegebene mit jedem theoretisch erzeugten Spektrum. Um die gefundenen Ergebnisse miteinander vergleichen zu können, war es nötig, eine Funktion zu implementieren, die es gestattet, sowohl die Quantität als auch die Qualität der gefundenen Treffer zu berücksichtigen. Die Intensität der Peaks findet keine Berücksichtigung, da es mit dem Programm *GLYCO-FRAGMENT* nicht möglich ist, Intensitäten vorherzusagen. Die Berechnung des quantitativen Anteils dieser Funktion war relativ einfach. Es wurde einfach die Anzahl der im Fehlerbereich gefundenen Peaks zu den eingegebenen Peaks ins Verhältnis gesetzt. Die Berechnung des qualitativen Anteils musste den Fehlerbereich mit berücksichtigen, da eine Abweichung von 0,1 Da bei einem Fehlerbereich von 2,0 Da bedeutend genauer ist als bei einem Fehlerbreich von 0,2 Da. Dies wurde erreicht indem der Quotient der Differenz von gemessenem Peak und berechnetem Peak und des Fehlerbereiches von Eins abgezogen wird. Dies bedeutet, dass ein Wert von Eins einem idealen Treffer entspricht, und ein Wert von 0,5 bedeutet, dass sich der gemessene Peak in der Mitte zwischen Peak und Fehlergrenze befindet. Diese Überlegungen führten zu der folgenden Formel:

$$MS_{SCORE} = \frac{\sum_{1}^{n} \left(1 - \frac{|P_s - P_r|}{Err}\right)}{n_{input}} * 100$$

Abbildung 64: Bewertungsfunktion des Suchalgorithmus

n	entspricht der Anzahl der eingegebenen Peaks
P_s	entspricht dem gemessenen Peak
P_r	entspricht dem berechneten Peak in der Bibliothek
Err	entspricht dem eingegebenen Fehlerbereich

Diese Funktion liefert einen Score zwischen Null und Einhundert, wobei ein Score von Einhundert bedeutet, dass die beiden Spektren identisch sind und ein Score von Null, dass keinerlei Gemeinsamkeiten zwischen diesen beiden Spektren existieren.

4.14.3 Das Webinterface

Zur Interaktion mit dem Benutzer war es nötig, ein Interface zu entwickeln, dass die Eingabe der erforderlichen Messparameter und der Peakliste ermöglichte:

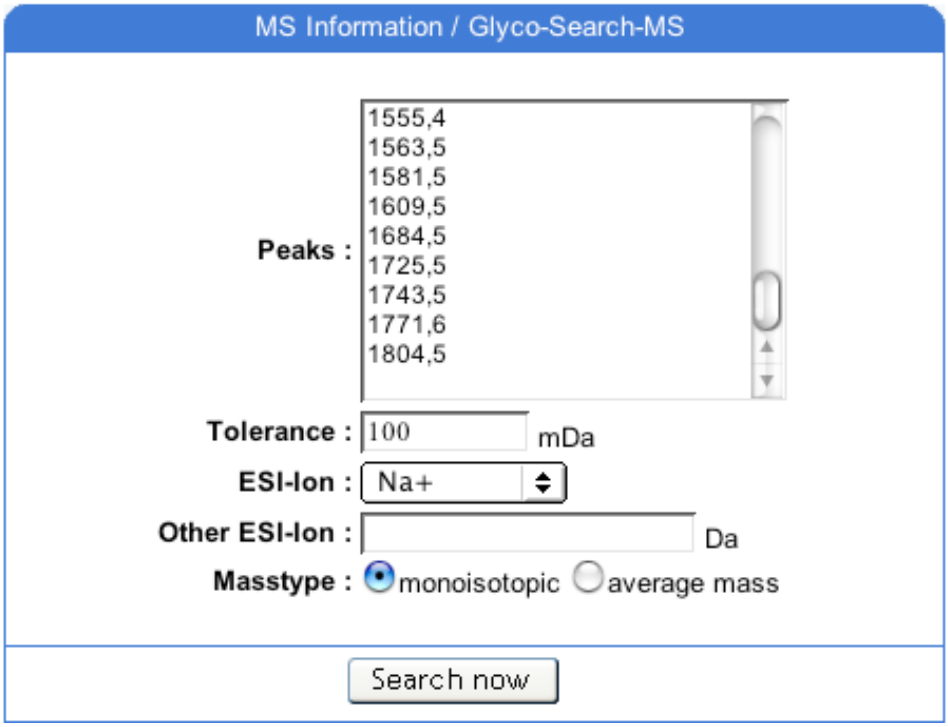


Abbildung 65: Eingabemöglichkeiten des Web-Formulars

Nach Eingabe der benötigten Parameter kann der Benutzer durch Anklicken des <Search now>-Buttons in der Spektral-Bibliothek suchen lassen. Das Durchsuchen der 11000 Spektren dauert dabei in der Regel etwa 10 Sekunden. Danach werden die n besten Treffer angezeigt.

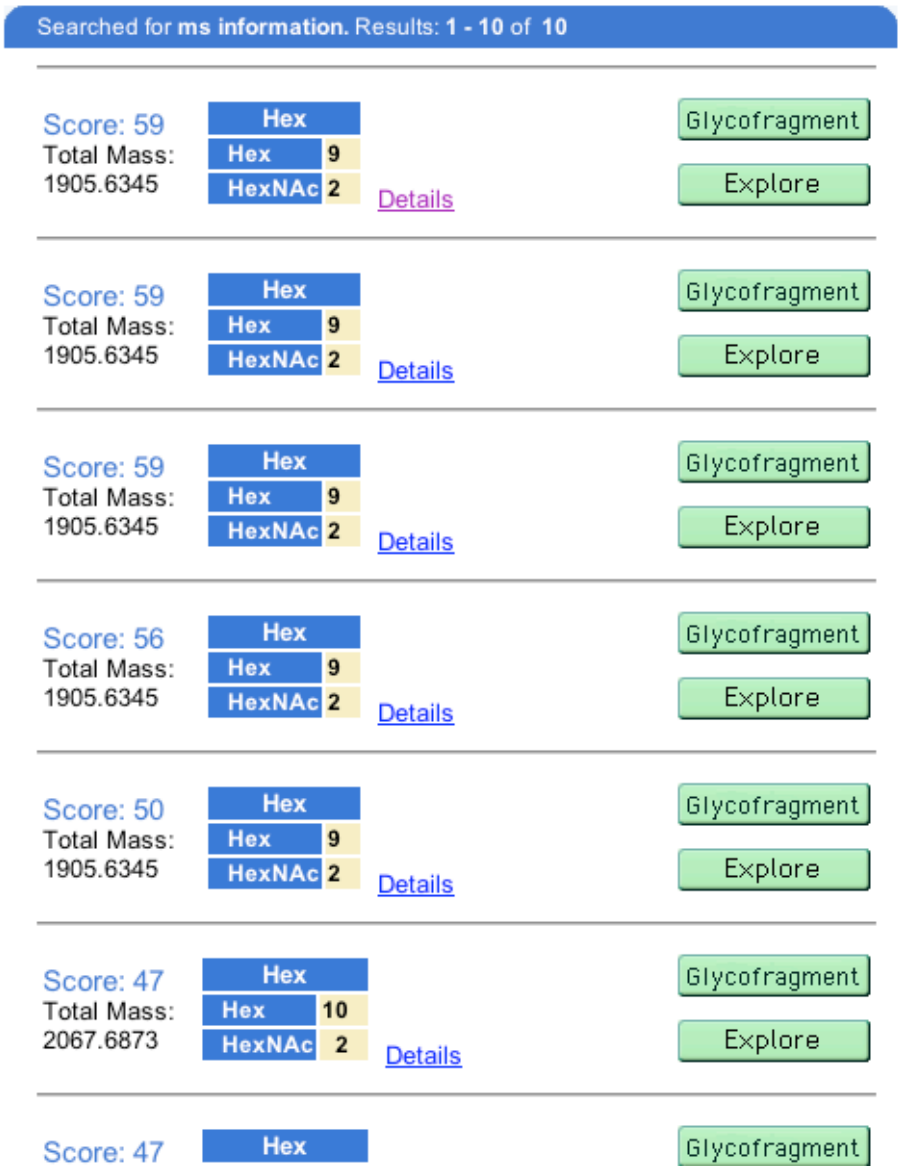


Abbildung 66: Die Liste der gefundenen Strukturen

Von hier aus kann dann eine genauere Untersuchung der gesuchten Verbindung mit dem Programm *GLYCO-FRAGMENT* (Anklicken des <Glycofragment-Buttons>) erfolgen oder man kann sich durch Klicken des Buttons <Explore> die weiteren in der *SWEET-DB* gespeicherten Daten anzeigen lassen. Für einen ersten Überblick der gefundenen Strukturen sind die Masse und die Komposition der Verbindungen angegeben. Weitere Einzelheiten des Suchergebnisses können in einer Detailansicht (Anklicken des <Details>-Links) angesehen werden.

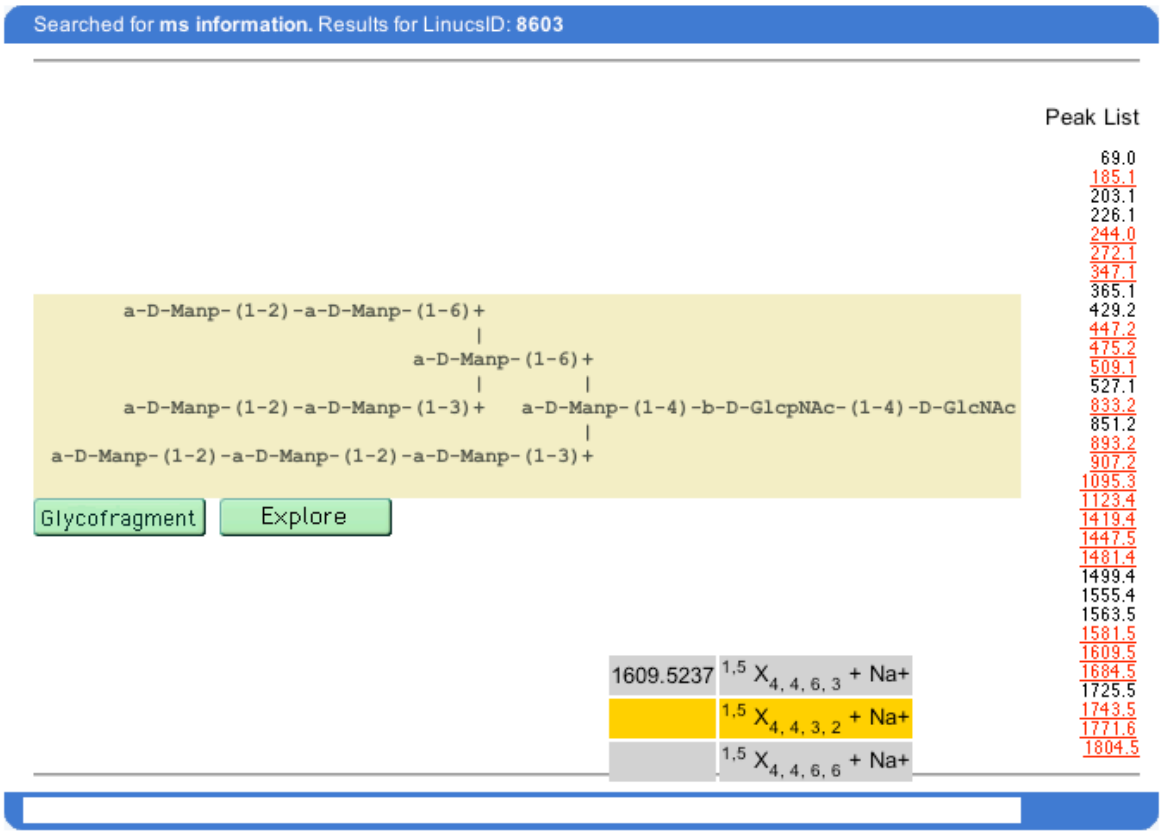


Abbildung 67: Die Detailansicht einer gefundenen Struktur

In der Detailansicht wird zum einen die gefundene Struktur dargestellt, zum anderen aber auch die eingegebene Peak-Liste. Die Treffer werden in rot dargestellt und können mit der Maus angeklickt werden, um in einer Tabelle die genauen Massen und die Trefferbezeichnungen anzuzeigen. Bei den nicht getroffenen Massen kann es sich um Ionen, die sich nicht nach den Regeln von Domon und Costello bilden, handeln. Es können allerdings auch interne Fragment-Ionen oder Ionen sein, die mehrfach geladen oder sich durch ein anderes als das ausgewählte Esi-Ion bilden.

4.14.4 Beispiele

In den folgenden drei Beispielen soll gezeigt werden, wo die Grenzen dieser Bibliothek liegen, aber auch welche Möglichkeiten der Suche in dieser Datenbank möglich sind. Dazu wird ein Spektrum von einem N-Glykan mit den Spektren in der Bibliothek verglichen. Das nächste Beispiel beschreibt die Suche eines derivatiserten N-Glykans. Das letzte Beispiel zeigt die Suche nach einem Lipopolysaccharid.

4.14.4.1 Suche nach einem N-Glykan

Die Substanzklasse der N-Glykane ist sehr gut von der Gruppe um Raymond Dwek am Glycobiology Institute in Oxford untersucht worden. Dort hat sich David Harvey im Bereich der Massenspektrometrie sehr intensiv mit der Analyse von N-Glykanen beschäftigt[54, 57, 89, 93-95]. Um den Suchalgorithmus und die Bewertungsfunktion zu überprüfen, wurde nun aus diesen Publikationen eine Struktur ausgewählt und das zugehörige Spektrum abgetippt. So wurde für einen ersten Test die folgende Struktur genommen:

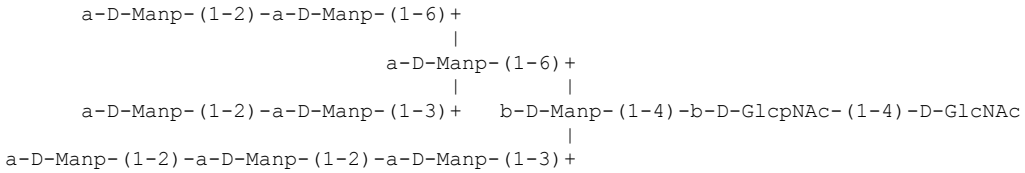
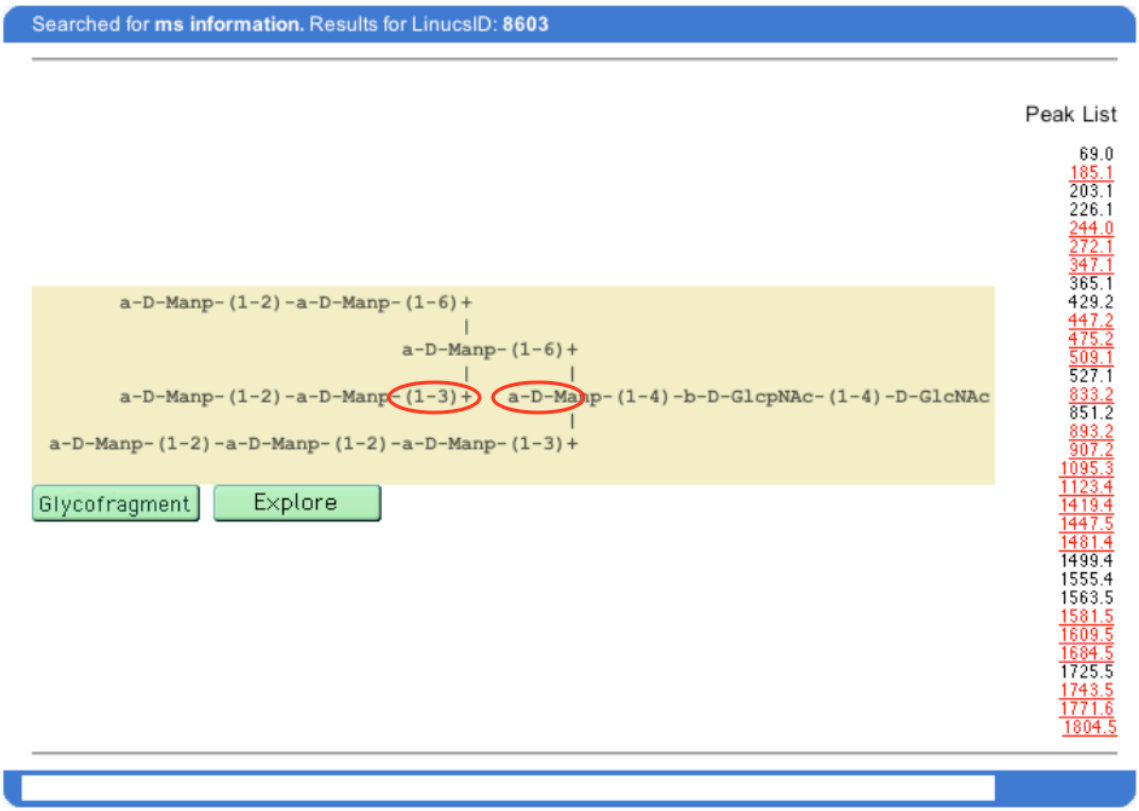


Abbildung 68: gesuchte Struktur

Das dazu in der Publikation befindliche Spektrum enthielt die folgende Peak-Liste:

69,0; 185,1; 203,1; 226,1; 244,0; 272,1; 347,1; 365,1; 429,2; 447,2; 475,2; 509,1; 527,1; 833,2; 851,2; 893,2; 907,2; 1095,3; 1123,4; 1419,4; 1447,5; 1481,4; 1499,4; 1555,4; 1563,5; 1581,5; 1609,5; 1684,5; 1725,5; 1743,5; 1771,6; 1804,5

Das Spektrum enthielt monoisotopische Massen und als Elektrospray-Ion war Na^+ verwendet worden. Die erlaubte Abweichung, um einen Peak als Treffer zu erkennen, betrug 100mDa. Mit dieser Peakliste und den drei Parametern wurde nun innerhalb der Spektren-Datenbank gesucht. Nach einer Suchzeit von weniger als drei Sekunden erhielt man die Trefferliste. Die drei besten Treffer hatten einen Score von 59 und eine Masse von 1905,6345Da, die der real gemessenen Masse von 1905,6 sehr nahe kam. Zu diesen drei Treffern gehörten die folgenden Strukturen:



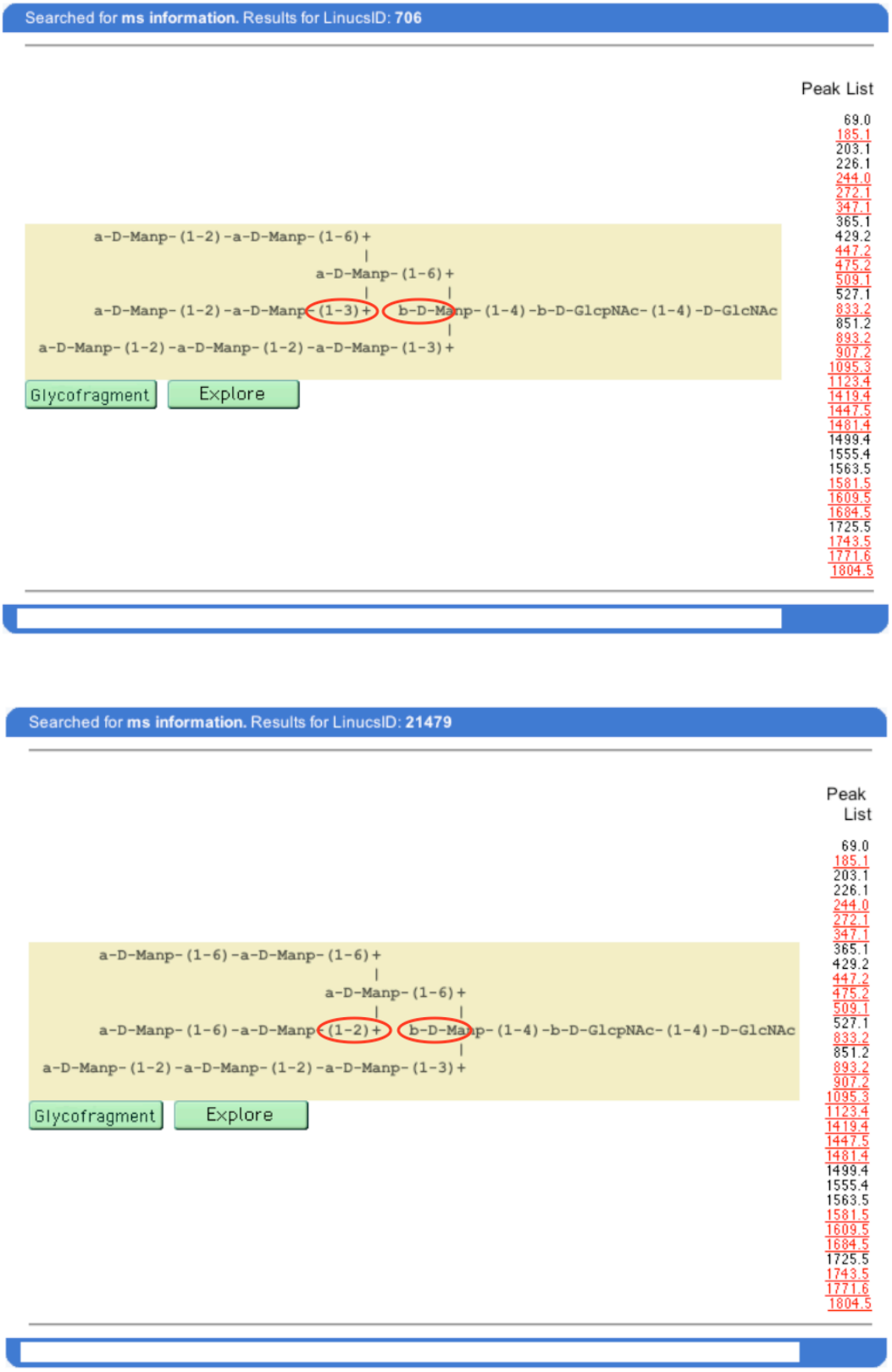


Abbildung 69: Ergebnisse der Suche

Diese Strukturen sehen auf den ersten Blick identisch aus, und bei allen wurden dieselben 21 von den eingegebenen 32 Peaks gefunden. Erst bei genauerem Hinsehen erkennt man, dass sich die ersten beiden Strukturen nur in der Stellung

einer Hydroxyl-Gruppe in dem dritten Residuum unterscheiden. Bei der dritten Struktur sind alle Residuen an gleicher Stelle vorhanden. Sie unterscheiden sich nur durch die Art der glykosidischen Bindungen. Die gesuchte Struktur befindet sich in der Ergebnismenge und eine echte Unterscheidung der Strukturen ist mit dem gemessenen MS-Spektrum leider nicht möglich. Eine eindeutige Unterscheidung der ersten beiden Strukturen wäre nur mit Hilfe der NMR möglich. Bei einer Optimierung der Messbedingungen, so dass auch für das □-D-Manp Residuum mit der Linkage-Information 4,4,6 alle A- und X-Ionen entstehen, kann jedoch diese Struktur von den anderen abgegrenzt werden.

4.14.4.2 Beispielsuche nach einem derivatisierten N-Glykan

Natürlich sollte auch gezeigt werden, dass die Suche auch für andere Substanzklassen als N-Glykane möglich ist. Dank Frau Geier aus Giessen standen hier einige real gemessene Spektren von derivatisierten Glykanen zur Verfügung[96]. Diese Spektren wurden dann als Eingabe für den Suchalgorithmus benutzt. Dazu wurde nun ein typisches N-Glykan, das am reduzierenden Ende mit 2-Aminopyridin derivatisiert ist, ausgewählt.

```
b-D-Galp-(1-6)+
              |
         a-D-Manp-(1-6)+              a-L-Fucp-(1-6)+
                       |                            |
                  b-D-Manp-(1-4)-b-D-GlcpNac-(1-4)-HexNac-2Ap
```

Das dazugehörige Massenspektrum bestand aus den folgenden Werten:

203,80; 281,91; 300,07; 301,04; 330,96; 413,12; 417,62; 424,99; 425,40; 446,21; 446,93; 467,80; 496,97; 497,53; 497,91; 498,42; 505,95; 506,42; 561,42; 569,93; 570,48; 571,40; 572,61; 607,81; 608,75; 699,56; 700,32; 711,83; 712,89; 826,88; 832,85; 834,05; 848,89; 849,84; 850,12; 994,98; 995,94; 1010,94; 1011,91

Die Suche mit diesen Werten lieferte für die richtige Struktur den höchsten Score von 53.

4.14.4.3 Beispielsuche nach einem Lipopolysacharid

Als letztes Beispiel wurde versucht, ob es auch möglich ist, ein Spektrum von einem Lipopolysaccharid aufzuklären. Hier wurde wieder auf ein Beispiel aus der Literatur zurückgegriffen, da leider keine gemessenen Spektren zur Verfügung standen. Dazu wurde auf eine Publikation aus dem Arbeitskreis von Prof. Peter-Katalinic aus Münster zurückgegriffen. Dabei wurde als Struktur das Gangliosid GM1 ausgewählt.

```
b-D-Galp-(1-3)-b-D-GalNAc-(1-4)-b-D-Galp-(1-4)-b-D-Glcp-(1-1)-Cer1618
                                              |
                    a-D-Neup5Nac-(2-3)+
```

Hier wurde auch zum ersten Mal ein Spektrum ausgewertet, das im negativen Modus gemessen worden ist. Die dazugehörige Peakliste bestand aus den folgenden Massenwerten:

290,1; 364,1; 408,1; 424,1; 564,6; 611,2; 655,3; 726,6; 768,6; 835,3; 888,7; 983,6; 1025,7; 1091,7; 1135,8; 1179,8; 1253,8; 1322,8; 1382,8; 1484,8; 1500,9; 1526,8; 1544,8

Die Suche mit diesen Werten lieferte für die richtige Struktur den höchsten Score von 46.

4.15 GLYKAN-PROFILING

Häufig wird bei der Auftrennung eines Kohlenhydratgemisches nicht ein komplettes Massenspektrum aufgenommen, sondern es kommt nur zur Bestimmung des Massenpeaks der getrennten Kohlenhydrate[97, 98]. Dieses geschieht dann für alle Bestandteile des Gemisches und nennt sich Glykan-Profiling. Gerade dafür ist es von essentieller Bedeutung, mittels geeigneter Spektren-Bibliotheken, eine schnelle Identifizierung der analysierten Zuckerstruktur zu ermöglichen, um eventuelle Rückschlüsse auf das in der untersuchten Zelle oder in dem untersuchten Gewebe veränderte Kohlenhydratgemisch ziehen zu können. Es wurde daher ein Webinterface entwickelt, das es gestattet, nur nach einer einzelnen Masse suchen zu können.

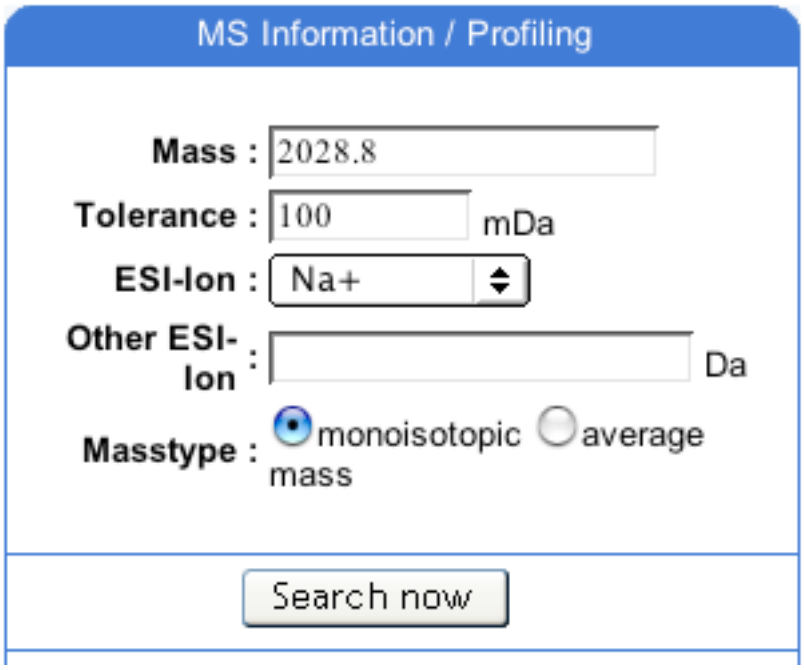


Abbildung 70: Webinterface des Glykan-Profiling Algorithmus

Durch Anklicken des <Search now>-Buttons wird nun in der Bibliothek gesucht. Es wird aber nur die molekulare Masse der Verbindung berücksichtigt. Auch hier wird nun eine Liste ausgegeben, so dass die besten Treffer oben stehen.

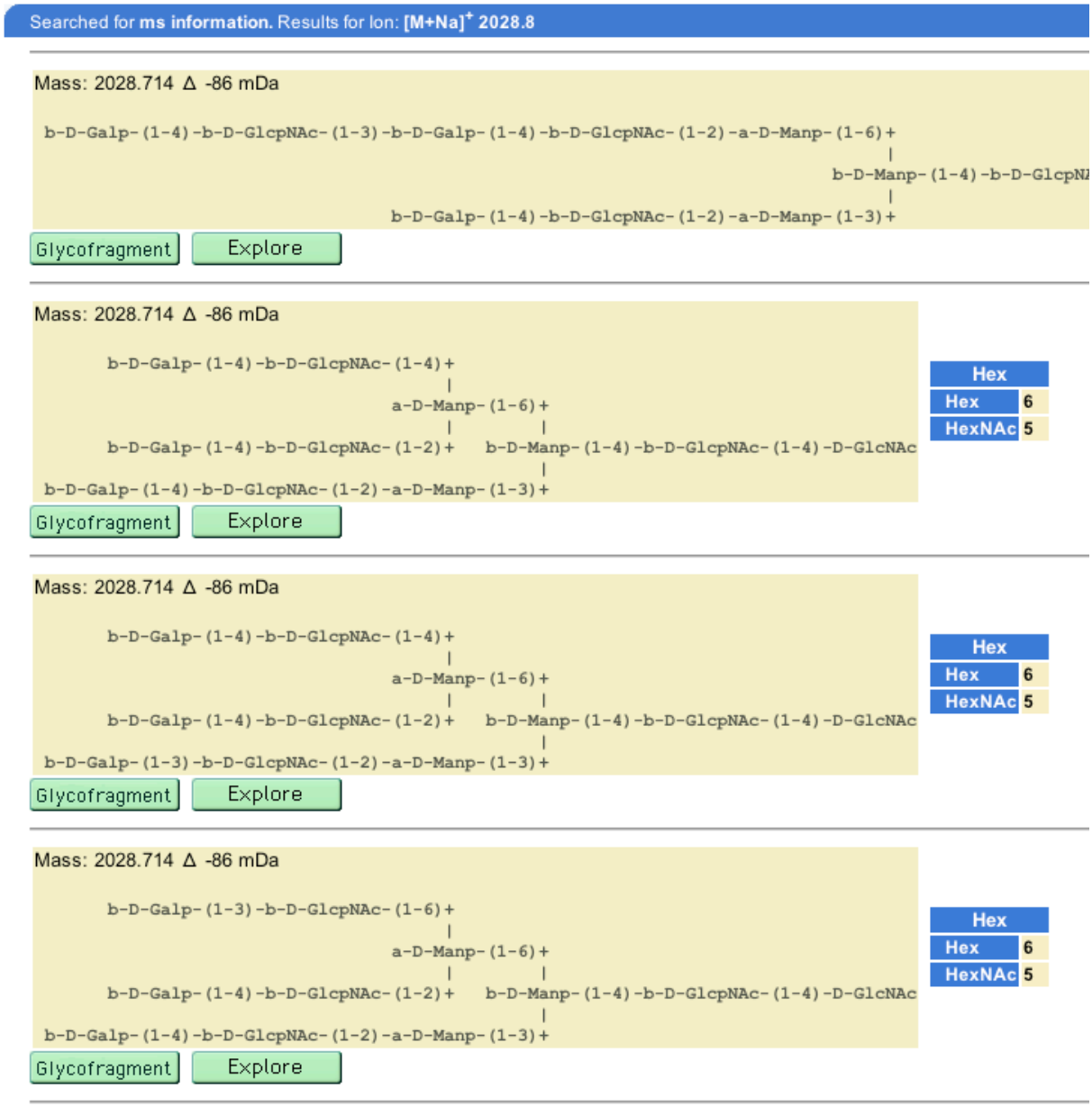


Abbildung 71: Liste der gefundenen Strukturen

Auch hier ist wieder gut zu erkennen, dass sich die besten Treffer in der Struktur sehr ähnlich sind. Dies bedeutet aber auch, dass eine eindeutige Zuordnung mit dieser Methode nicht möglich ist. Wenn in einer zukünftigen Version noch weitere Parameter wie Spezies oder Gewebe hinzukommen, kann aber auch auf diese Art und Weise eine Identifizierung möglich sein.

4.15.1 Ergebnis und Diskussion

Mit den beiden vorgestellten Methoden *GLYCO-SEARCH-MS* und *GLYKAN-PROFILING* bietet sich zum ersten Mal die Möglichkeit, Kohlenhydrat-Strukturen, entweder durch den ermittelten Molekülpeak oder durch ein gemessenes MS-Spektrum, automatisch an Hand von theoretischen Vergleichsspektren zu ermitteln. Im Gegensatz zu den Programmen *GLYCOMOD* und *GLYCANMASS*, die in der Einleitung beschrieben sind, finden hier zum ersten Mal auch die strukturell aussagekräftigeren X- und A-Ionen der Kohlenhydrate[54] bei der Ermittlung der Struktur Berücksichtigung. Auf Grund der bedeutend größeren Residuenanzahl (für

GLYCOMOD und *GLYCANMASS* nur 10, *GLYCO-SEARCH-MS* und *GLYKAN-PROFILING* 450), die für die Beschreibung der Strukturen zur Verfügung stehen, können auch wesentlich mehr Fragestellungen bearbeitet werden. So können sowohl N- und O-Glykane, sowie deren Derivate, als auch Lipopolysaccharide mit dem Suchalgorithmus erkannt werden. Ideal ist das Verfahren für einen automatischen Vergleich bei der Auftrennung mittels chromatographischer Verfahren. Dazu muss die Bewertungsfunktion aber noch weiter optimiert werden, so dass eine eindeutige Identifizierung des Kohlenhydrats möglich ist. Zurzeit können Scores im Bereich von 40 bis 75 als gut zugeordnet betrachtet werden, aber leider nicht als erkannt gelten. Zumindest aber die Komposition ist richtig zugeordnet. Ab einem Score von 75 kann die Substanz als erkannt gelten. Dieses hängt aber noch sehr stark von der Qualität des ermittelten Massenspektrums ab, da in der Literatur bei vielen Massenspektren die Auflösung noch zwischen 0,5 und 1,0 Da liegt. Bei MS-Spektren deren Auflösung unter 0,1 Da liegt, und deren Messparameter optimal für Kohlenhydrate sind, so dass viele aussagekräftige A- und X-Fragmente entstehen, kann aber eine eindeutige Zuordnung der Komposition und auch der glykosidischen Bindungen erfolgen[89, 93]. Für die Zukunft ideal wäre eine Suche, bei der auch taxonomische Parameter Berücksichtigung finden, damit eine große Anzahl von Strukturen von vornherein ausgeschlossen werden können.

5 Untersuchungen zur automatischen Pflege und Annotierung einer Datenbank

Das Internet hat grundsätzlich und praktisch die Art und Weise verändert, wie wissenschaftlichen Daten, Informationen und Wissen ausgetauscht werden. Mit dem Internet sind zum ersten Mal die technischen Möglichkeiten geschaffen, Wissen weltweit zu präsentieren und allgemein zugänglichen zu machen. Internet basierte Datenbanken sind ein wesentlicher Bestandteil der wissenschaftlichen Informationsvermittlung im Bereich der Molekularbiologie und Bioinformatik. Sie bieten einen einfachen und standardisierten Zugang über Webbrowser von jedem Punkt der Welt, der über einen Internetanschluss verfügt.
Im Zuge der Sequenzierung des menschlichen Genoms haben sich für den Bereich der Genomik und Proteomik in den letzten 10 Jahren sehr vielfältige Datenbanken etabliert, die eine komfortable Recherche und Extraktion von Daten unter verschiedenen medizinischen und biologischen Aspekten gestatten. Für die Genomik sind dies in erster Linie die Datenbanken *GENBANK* und *EMBL*. Für den Bereich der Proteomik sind die Datenbanken *SWISSPROT*/*TREMBL* und die *PROSITE* zu nennen. Die Anstrengungen zum Aufbau von Datenbanken im Bereich der Glykobiologie nehmen sich dagegen eher bescheiden aus. Allerdings gibt es in den letzten zwei Jahren verschiedene Initiativen, entsprechende Datenbanken auch für die Glykowissenschaften zu entwickeln. Als Beispiele sind hier zu nennen die CAZy-(Carbohydrate Active Enzymes) Datenbank, die *O-GLYCBASE* (experimentell bestätigte O-Glykosylierungspositionen), die *GLYCOSUITEDB* (N- und O-Glykane) und die im DKFZ entwickelte *SWEET-DB* als wichtigste Vertreter zu nennen.
Ein wichtiger Faktor für die Akzeptanz einer wissenschaftlichen Datenbank ist ihre Aktualität, ein einfacher und intuitiver Zugang und die Qualität, der in ihr enthaltenen Daten. Die Organisierung des Prozesses der Neueingabe von wissenschaftlichen Daten ist daher ein Prozess von zentraler Bedeutung, der leider auch sehr arbeits- und damit auch kostenintensiv ist. So konnten die frühen großen Datenbankprojekte im Bereich der Glykowissenschaften, *CARBBANK* und *SUGABASE*, nicht mehr aktualisiert werden, nachdem die staatlichen Förderungen ausgelaufen waren.
Bei der *CARBBANK* und *SUGABASE* wurde die Neuaufnahme von Strukturen und deren Annotation weitgehend manuell vorgenommen. Dies ist jedoch ein sehr empfindlicher Prozess für eine Datenbank. So wird sie für den Benutzer unbrauchbar, sobald sich zu viele fehlerhafte oder nicht in das Themengebiet gehörende Daten darin befinden. Die notwendige Kontrolle einer manuellen Eingabe erhöht jedoch den Zeitbedarf und die Kosten der Datenbank erheblich. Allerdings sind viele Daten, die früher aus wissenschaftlichen Publikationen durch einen Wissenschaftler extrahiert und manuell eingegeben werden mussten, heute in digitaler Form verfügbar. Mit der *PUBMED* steht ein für die Biowissenschaften umfassendes Reservoir an wissenschaftlichen Erkenntnissen allgemein in Form von Abstracts zur Verfügung. Zusätzlich können weitere wissenschaftliche Daten in Form von digitalen Publikationen erschlossen werden, wenn, wie dies im DKFZ der Fall ist, für relevante Zeitschriften die Zugriffsrechte vorhanden sind. Vor diesem Hintergrund eröffnen sich neue Wege der automatischen Erschließung und Aufbereitung von wissenschaftlichen Daten. In den folgenden Abschnitten soll untersucht werden, inwieweit es möglich ist, spezielle wissenschaftliche Datensammlungen, hier die bereits existierende *SWEET-DB*, durch die Verwendung

von wissenschaftlichen Ansätzen aus den Informationswissenschaften zu komplettieren. Weiterhin soll untersucht werden, wieweit sich die Beschreibung der in der Literatur bekannten Informationen über das Vorkommen und die Funktionen von Glykanen durch den Einsatz von neuen Such- und Bewertungsalgorithmen automatisieren lässt.
Im Moment wird in Amerika durch das Consortium for Functional Glycomics[1] versucht, eine Datenbank mit Informationen zu Kohlenhydraten, wie Vorkommen in welcher Spezies, Struktur, chemische und physikalische Eigenschaften und deren physiologischen Funktionen zu schaffen. Im Rahmen dieser Entwicklung werden auch unterschiedliche Annotierungsstrategien untersucht.
Weiterhin gibt es viele kleinere Datenbankprojekte, von denen eine Auswahl in den folgenden Abschnitten dargestellt wird. Einen aktuellen Überblick über im Internet verfügbare biomedizinische Datenbanken liefert jährlich das erste Heft eines Jahrgang von Nucleic Acids Research „The Database Issue 2003“[65].
Die Verknüpfung der Gen- und Protein-Datenbanken über ihre jeweilige Sequenz hat sich für die effiziente Nutzung dieser Datenbanken als sehr wichtig erwiesen. Für ein bestimmtes Protein kann der Benutzer so schnell alle verfügbaren Informationen aus verschiedenen Datenbanken abrufen. Die Entwicklung von effizienten Ansätzen zur Quervernetzung von Informationen aus biomedizinischen Datenbanken, die unter verschiedenen wissenschaftlichen Aspekten erhoben wurden, hat sich als eine Technik erwiesen, die ein hohes Potential für synergetische Effekte in der biomedizinischen Forschung beinhaltet.

[1] http://web.mit.edu/glycomics/consortium/

5.1 Im Web verfügbare Datenbanken im Bereich der Glykobiologie

Name	Inhalte	URL
Kohlenhydrat-relevante Informationen in Protein Datenbanken		
CAZy	Kohlenhydrat-aktiven Enzyme	http://afmb.cnrs-mrs.fr/CAZY/
Lectines	3D Strukturen von Lektinen	http://www.cermav.cnrs.fr/lectines
CTDL	tierische Lektine	http://ctld.glycob.ox.ac.uk/
PDB2LINUCS	Glykoproteine in der PDB	http://www.dkfz.de/spec/pdb2linucs/
Räumliche Strukturen		
Disaccharides	Konformationskarten	http://www.cermav.cnrs.fr/cgi-bin/di/di.cgi
GlycoMaps DB	Konformationskarten	http://www.dkfz.de/spec/glycomaps/
Kohlenhydrat-Datenbanken		
SWEET-DB	DKFZ-Heidelberg	http://www.dkfz.de/spec/sweetdb/
Carbohydrate DB	Consortium for Functional Glycomics	http://web.mit.edu/glycomics/carb/carbdb.shtml
GlycoSuiteDB	Proteome Systems Ltd	http://www.glycosuite.com/
Glycomic Database	GlycoMinds	http://www.glycominds.com/GlycoInfo.asp

Tabelle 13: Im Web verfügbare Datenbanken im Bereich der Glykobiologie

5.2 Ausgangssituation

Im Laufe der vergangenen Jahre sind die oben genannten über das Internet zugänglichen Datenbanken entwickelt worden. Teilweise werden diese Datenbanken schon über mehrere Jahre intensiv gepflegt. Es wird von den Entwicklern der Datenbanken permanent daran gearbeitet, dass der Datenbestand der Datenbanken sich stetig erhöht und auch neue Möglichkeiten der Recherche oder andere Features hinzukommen. Im Folgenden sind einige der Datenbanken näher erläutert, die für den Bereich der Genomik, Proteomik und Glykomik als repräsentativ gelten können.

5.2.1 GENBANK

Bei der *GENBANK*[99][1] handelt es um eine Datenbank des National Institut of Health (NIH) der USA für alle Gen- und Proteinsequenzen, die öffentlich zugänglich sind. Derzeit sind mehr als zwölf Milliarden DNA-Bausteine in der Datenbank gespeichert. Diese Gensequenzen stammen von über 55 000 verschiedenen Organismen, darunter natürlich auch die Daten des Humangenoms. Alle Gensequenzen und Proteindaten werden direkt mit Hilfe definierter Schnittstellen von einzelnen Genforschern oder von Großprojekten eingeschickt. Tägliche Aktualisierungen können über das Internet über FTP-Server abgerufen werden. Per E-Mail können schließlich sämtliche Gen- und Proteindaten jederzeit über ein Interface abgefragt werden, ohne im Internet die Homepage der NIH ansteuern zu müssen. Die *GENBANK* enthält außer den Sequenzdaten auch Informationen zur Taxonomie, Kartierung, Proteinstruktur und biomedizinische Literaturangaben mit den Kurzfassungen der publizierten Studien. Die Datenbank kann kostenlos über die Homepage des National Center for Biotechnology Information (NCBI) am NIH durchforstet werden.

[1] http://www.ncbi.nlm.nih.gov

5.2.2 EMBL

Die *EMBL* Nucleotide Sequence Database[1], die vom European Bioinformatics Institute (EBI) in Hinxton zusammengestellt wird, enthält Informationen über entschlüsselte Nukleinsäuresequenzen. Die zurzeit aktuelle Version 76 vom September 2003 enthält über 27 Millionen Sequenzen mit insgesamt über 33 Milliarden Nukleotiden. Seit der Version 1 vom Jahre 1982 vermehrte sich der Inhalt der Datenbank um 50%-100% in jedem Jahr. *EMBL* ist in 18 Abteilungen aufgeteilt, die jeweils Sequenzen aus abgegrenzten Forschungsbereichen zusammenfassen sollen. In den meisten Fällen ergibt sich diese Abgrenzung durch die Taxonomie der Herkunftsorganismen.

5.2.3 SWISSPROT/TREMBL

SWISSPROT[2], eine Datenbank mit Informationen zu Proteinseqenzen, wird gemeinsam von dem Department of Medical Biochemistry der Universität Genf und dem EBI in Hinxton zusammengestellt. Die aktuelle Version 41 stammt vom Februar 2003 und enthält 122.564 Sequenzen mit insgesamt fast 45 Millionen Bausteinen. Alle Sequenzen der aktuellen Version sind in einer Datei zusammengestellt. Es gibt also keine verschiedenen Abteilungen wie bei der *EMBL*. Die *TREMBL*-Datenbank entspricht den Übersetzungen aller kodierenden DNA-Sequenzen aus der *EMBL*-Datenbank, die noch nicht in der *SWISSPROT* enthalten sind.

5.2.4 PROSITE

Bei der *PROSITE* Datenbank[100-103][3] handelt es sich um den Versuch die Vielzahl der verschiedenen Proteine in Proteinfamilien zusammenzufassen. In den Proteinen können außerdem Domänen identifiziert werden. Auf die Einträge der *PROSITE* kann durch unterschiedliche Suchfunktionen wie Literaturstelle, Autor oder Beschreibung zugegriffen werden. Im Moment (Stand September 2003) sind in der Datenbank 1655 verschiedene Einträge enthalten.

5.2.5 Brookhaven Protein Databank (PDB)

Bei der Brookhaven Protein Databank[62, 103][4] handelt es sich um eine Datenbank, die die 3D-Koordinaten von Proteinen beinhaltet. Bei den 3D-Koordinaten handelt es sich entweder um Röntgenstrukturen oder durch Molecular Modelling berechnete Modelle.
Der Benutzer ist so in der Lage, nach einer speziellen Proteinstruktur für eine Spezies zu suchen. Diese Strukturen können dann z. B. für Docking-Experimente benutzt werden. Die Strukturen können in unterschiedlichen Formaten herunter geladen werden. Dies ist zum einen das pdb-format, das aber immer mehr durch das mmCIF-Format abgelöst wird.

5.2.6 PUBMED

Eine der wichtigsten und in dieser Arbeit am häufigsten benutzten Datenbank ist die *PUBMED*[5], eine ebenfalls vom NCBI gepflegte Datenbank, die Publikationsdaten für

[1] http://www.ebi.ac.uk/embl/index.html
[2] http://www.expasy.org
[3] http://us.expasy.org/prosite/
[4] http://www.pdb.org/
[5] http://www.pubmed.org

den Bereich der Medizin und Biotechnologie bereithält. Hier zeigt sich sehr schön, wie einfach und auch effektiv es sein kann, Millionen von Datensätzen zu durchsuchen und die Daten nach der Suche zu übernehmen. In den letzten Jahren hat die NCBI auch der allgemeinen Entwicklung im Internet Rechnung getragen und bietet seine Daten jetzt auch in einer objektorientierten XML-Form an, die direkt von einer Anwendung weiterverarbeitet werden kann. So werden durch die *ENTREZ PROGRAMMING UTILITIES* Schnittstellen zur Verfügung gestellt, die es dem Benutzer gestatten, auch ohne Webinterface Suchanfragen zu stellen und die Daten an das Programm zur Weiterverarbeitung übertragen. Die *PUBMED* enthält zur Zeit über 14 Millionen Einträge, die bis in das Jahr 1950 zurückreichen. In den letzten Jahren sind im Durchschnitt etwa 500.000 neue Publikationen pro Jahr hinzugekommen.

5.2.7 CAZy – Carbohydrate Active EnZymes

Bei der CAZy [9, 10][1] handelt es sich um eine Datenbank, die eine sehr detaillierte Übersicht der einzelnen Familien der kohlenhydrataktiven Enzyme enthält. Eine grobe Einteilung der Enzyme erfolgt in Glykosidasen, Transglykosidasen, Glykosyltransferasen, Polysaccharid Lyasen und Kohlenhydrat Esterasen. Es besteht eine Suchmöglichkeit nach Organismus oder Familienzugehörigkeit des Enzyms. Es wird, wenn bekannt, sowohl das Substrat der katalysierten Reaktion dargestellt, als auch das Endprodukt. Falls möglich erfolgt eine direkte Verlinkung zur *SWISSPROT*, *GENBANK* oder zur PDB-3D-Struktur.

5.2.8 O-Glycbase

In der O-GLYCBASE[104, 105][2] befinden sich Daten zu den O-Glykanen und Glykosylierungsstellen von Proteinen. Zurzeit (Stand September/2003) befinden sich 242 Einträge in der Datenbank. In den Datensätzen werden die Sequenz des Proteins, die Stelle der Modifikation und die Publikation, aus der die Daten entnommen worden sind, dargestellt. Gleichzeitig bietet die Website die Möglichkeit, sich Glykosylierungsstellen durch das Programm *NETOGLYC*[106, 107] vorhersagen zu lassen.

5.2.9 GLYCOSUITEDB

Bei der *GLYCOSUITEDB*[108, 109][3] handelt es sich um eine kommerzielle Datenbank von der Firma Proteome Systems Ltd., die Daten enthält, die in den Bereich der Glykobiologie gehören. Es befinden sich zurzeit 8100 verschiedene Strukturen in der Datenbank, wobei es sich um N-Glykane, O-Glykane und Lipopolysaccharide handelt. Die Suchmöglichkeiten sind sehr vielfältig. So kann nach Massen, beliebigen Substrukturen, Spezies, Publikationsdaten und auch der Komposition in der Datenbank gesucht werden. Es werden allerdings keine NMR-Spektren oder Massenspektren bereitgehalten, so dass eine Identifizierung von unbekannten Substanzen durch diese Methode nicht ohne weiteres möglich ist. Leider werden die Informationen seit März 2003 auch den akademischen Nutzern nicht mehr kostenlos zur Verfügung gestellt.

---

[1] http://afmb.cnrs-mrs.fr/CAZY/

[2] http://www.cbs.dtu.dk/databases/OGLYCBASE/

[3] http://www.glycosuite.com/

5.2.10 SWEET-DB

Ähnlich wie bei der *CARBBANK* wurde mit der *SWEET-DB*[59] [1] eine Datenbank entwickelt, die es zum Ziel hat, eine Arbeitsumgebung für alle Fragestellungen im Bereich der Kohlenhydratanalytik zu sein. Die Datenbank enthält die Daten der ehemaligen *COMPLEX CARBOHYDRATE DATABASE*[110] und der ehemaligen *SUGABASE*. Beide Datenbanken wurden leider wegen fehlender Finanzmittel und der aufwendigen Annotation der enthaltenen Daten eingestellt.
Im Laufe der Jahre wurden immer mehr Daten zur *SWEET-DB* hinzugefügt. Ein großer Vorteil der *SWEET-DB* ist die Verlinkung mit weiteren bestehenden Datenbanken: So besteht die Möglichkeit direkt zur Webseite der *PUBMED* zu springen, um von dort aus weitere Literaturrecherchen zu einem bestimmten Kohlenhydrat oder aber auch zu einem bestimmten Wissensgebiet zu machen. Großer Wert wurde auf die Abfragemöglichkeiten gelegt, mit denen es möglich ist, Daten aus der Datenbank zu selektieren. So besteht zum einen die Möglichkeit in den Titeln, aber auch in den Autoren der gespeicherten Publikationen nach einem bestimmten Schlagwort oder einem bestimmten Autor zu suchen. Bei beiden Suchmethoden kann man eine Ähnlichkeitssuche wählen, so dass auch ähnliche Begriffe oder Autoren gefunden werden.
Ein großer Vorteil besteht darin, dass die Strukturen intern in der so genannten *LINUCS*-Notation[111] gespeichert werden. Mit dieser Notation ist es möglich, die Sequenz eines Kohlenhydrats linear zu beschreiben, und man verbessert so die Computer-Lesbarkeit. Man kann so nach Substrukturen suchen, aber viel wichtiger ist die Möglichkeit Daten basierend auf der Struktur in der Datenbank zu finden. So kann schnell die Masse und auch die Summenformel und Komposition der Struktur ermittelt werden.
In der Datenbank befinden sich zurzeit 22337 Strukturen, die aus 15364 Publikationen extrahiert worden sind (Stand September 2003).

5.3 Annotierungsstrategien

Bei den meisten der Datenbanken erfolgt die Annotierung der Datenbank manuell, d. h. es existieren Schnittstellen, die es dem Benutzer gestatten, Daten in eine temporäre Datenbank einzutragen. Dieses Vorgehen wird bei der *GENBANK*, der PDB-Datenbank und der *SWISSPROT*/*TREMBL* Datenbank verfolgt. Diese Daten werden dann durch einen Experten gesichtet und wenn möglich mit weiteren Informationen versehen. Der Experte hat die Möglichkeit diesen Eintrag zurückzuweisen, weitere Experimente durchführen zu lassen oder den Eintrag in die eigentliche Datenbank zu überführen. Damit steht der Eintrag der Allgemeinheit zur Verfügung. Bei der CAZy-Datenbank werden sogar alle Informationen manuell aus der Literatur extrahiert. Dieser zentrale Annotierungsprozess ist natürlich sehr zeit- und personalintensiv, und hat in der Vergangenheit dazu geführt, dass Datenbankprojekte mangels finanzieller Ressourcen eingestellt worden sind. So ist dieses Geschehen bei der Carbbank[112] und auch der Sugabase[60, 61]. In den letzten Jahren sind verstärkt Anstrengungen unternommen worden, den Prozess der Datenextraktion zu automatisieren. So wird versucht die *SWISSPROT* durch die automatische Annotation von mikrobiellen Proteomen zu erweitern[113].
Eine weitere Möglichkeit der Datenbank Daten hinzuzufügen, besteht in der Möglichkeit automatisch Daten aus den Rohdaten oder bereits gespeicherten Daten

[1] http://www.dkfz-heidelberg.de/spec/sweetdb/

abzuleiten. Dieses umfasst auf der einen Seite so triviale Dinge wie das Berechnen der monoisotopischen und durchschnittlichen Molekularmasse einer gespeicherten Verbindung. Zum anderen können aber auch anspruchsvollere Dinge aus der Strukturformel abgeleitet werden, dieses kann die dreidimensionale Struktur, aber auch das theoretische Massenspektrum oder auch der Molekülpeak eines Massenspektrums sein. Ein gutes Beispiel ist die Übersetzung der DNA-Sequenzen aus der *EMBL*-Datenbank und Aufnahme in die *TREMBL*-Datenbank.

5.4 Aufgabenstellung

Seit dem Jahre 1999 sind keine neuen Daten zu den Datenbeständen der *CARBBANK* und *SUGABASE* hinzugefügt worden. Im Rahmen meiner Dissertation sollte nun untersucht werden, wie der Datenbestand einer bestehenden Datenbank, der *SWEET-DB*, durch den Einsatz bioinformatorischer Ansätze möglichst einfach erweitert werden kann. Dazu sollten Schnittstellen definiert werden, die es gestatten, durch ein klar definiertes Protokoll Daten in der Datenbank zu verwalten. Die Updatestrategien sollten unter Berücksichtigung der folgenden Schwerpunkte entwickelt werden:

- Effiziente Neueingabe und Verwaltung der vorhandenen Daten
- Extraktionsmöglichkeiten und Quervernetzung zu bestehenden Datenbankprojekten
- Aus vorhandenen Daten sollten neue abgeleitet und berechnet werden
- Der Datenbestand sollte möglichst vollautomatisch aktuell gehalten werden. Dazu sollte versucht werden, aus der Literatur automatisch neue Strukturdaten zu extrahieren und Texte automatisch in den Bereich der Glykobiologie einzuordnen.

Da aufgrund der Verwendung von vielen unscharfen Begriffen in der wissenschaftlichen Literatur nicht zu erwarten ist, dass die nachträgliche Extraktion von Daten, Informationen und Wissen immer erfolgreich sein wird, ist es ein weiteres Ziel, die Ansätze der automatischen Extraktion soweit voranzutreiben, dass dem eingebenden Experten möglichst umfassende Hilfestellungen gegeben werden. Sollte es nicht möglich sein, diese Daten vollautomatisch in die Datenbank zu schreiben, sollte versucht werden, den Aufwand bei der Selektion für den Experten möglichst gering zu halten. Dieses betrifft in erster Linie die Sichtung von automatisch extrahierten Daten, aber auch die Bereitstellung entsprechender Tools, die es einem Student ermöglichen, fehlerfrei Daten aus frei zugänglichen Quellen einzugeben.

5.5 Eigene Arbeiten

Zuerst wurde eine Schnittstelle zum Verwalten von Literaturstellen definiert. Diese ist beispielhaft für die Datenverwaltung der *SWEET-DB* und kann auch für andere Bereiche der Datenbank übernommen werden. Dann wurde begonnen, aufbauend auf diese Schnittstellen Webinterfaces und Applikationen zu entwickeln, die entweder der

- Eingabe von Daten durch Studenten dienen
- Einem Experten eine Hilfestellung bei der Selektion von automatisch extrahierten Daten geben.

Außerdem wurde die *SWEET-DB* um eine theoretische Spektral-Datenbank für Massenspektren erweitert.

5.6 Definition einer Schnittstelle zur Pflege der Publikationsdaten der *SWEET-DB*

Bis jetzt waren noch keine Schnittstellen entwickelt worden, die ein einfaches Verwalten von Daten in der *SWEET-DB* ermöglichen. Diese Verwaltung muss einfach, aber auch sehr sicher sein, damit es zu keinen Datenverlusten oder Inkonsistenzen durch den Benutzer kommen kann. Grundsätzlich müssen durch diese Schnittstelle drei verschiedene Funktionen gewährleistet sein:

1. Anzeige der Daten
2. Neueintragen von Daten
3. Löschen von Daten

In dieser Arbeit wurden sowohl Programme verwendet, die auf den Programmiersprachen ‚C' und ‚PHP' basieren. Dieses bedeutet, dass für beide Programmiersprachen Schnittstellen entwickelt werden mussten.

5.6.1 Anforderungen an die Schnittstelle

Zusätzlich zu den grundsätzlichen Anforderungen gab es natürlich weitere Regeln, die durch die Schnittstelle beachtet werden mussten. Die Daten der *SWEET-DB* werden im Moment in einer MySQL Datenbank gespeichert. Die dazugehörigen Tabellen enthalten dabei die folgenden Datenfelder:

```
+-------------+--------------+------+-----+---------+----------------+
| Field       | Type         | Null | Key | Default | Extra          |
+-------------+--------------+------+-----+---------+----------------+
| ReferenceID | int(11)      |      | PRI | NULL    | auto_increment |
| UserID      | int(11)      |      | MUL | 0       |                |
| StatusFlag  | int(11)      |      | MUL | 0       |                |
| InputID     | int(11)      |      | UNI | 0       |                |
| JournalName | varchar(254) |      | MUL |         |                |
| Volume      | varchar(10)  |      | MUL |         |                |
| Authors     | varchar(254) |      | MUL |         |                |
| Title       | varchar(254) |      | MUL |         |                |
| Year        | varchar(254) |      | MUL |         |                |
| Pages       | varchar(254) |      |     |         |                |
| PubmedID    | bigint(20)   |      |     | 0       |                |
+-------------+--------------+------+-----+---------+----------------+
```

Listing 4: Tabellenstruktur der Tabelle Reference

Zusätzliche Informationen zu der obigen Tabelle werden in der folgenden Tabelle ReferenceMore gespeichert.

```
+-----------------+---------+------+-----+---------+-------+
| Field           | Type    | Null | Key | Default | Extra |
+-----------------+---------+------+-----+---------+-------+
| ReferenceID     | int(11) |      | PRI | 0       |       |
| Keywords        | text    | YES  |     | NULL    |       |
| Keywords_Match1 | text    | YES  |     | NULL    |       |
| Keywords_Match2 | text    | YES  |     | NULL    |       |
| Abstract        | text    | YES  |     | NULL    |       |
| Abstract_Match1 | text    | YES  |     | NULL    |       |
| Abstract_Match2 | text    | YES  |     | NULL    |       |
+-----------------+---------+------+-----+---------+-------+
```

Listing 5: TabellenStruktur der Tabelle ReferenceMore

Die Informationen, welche Strukturen in der jeweiligen Publikation enthalten sind, werden in einer dritten Tabelle gespeichert.

```
+-------------+---------+------+-----+---------+-------+
| Field       | Type    | Null | Key | Default | Extra |
+-------------+---------+------+-----+---------+-------+
| ReferenceID | int(11) |      | PRI | 0       |       |
| LinucsID    | int(11) |      | PRI | 0       |       |
| UserID      | int(11) |      | MUL | 0       |       |
| StatusFlag  | int(11) |      | MUL | 0       |       |
| InputID     | int(11) |      | MUL | 0       |       |
+-------------+---------+------+-----+---------+-------+
```

Listing 6: Tabellenstruktur der Tabelle StructuresInReference

Diese Tabellen müssen durch Erstellen von konsisten MySQL-Queries richtig mit den eingegebenen Daten ausgefüllt werden.
Beim Neueintrag einer Publikation muss auf verschiedene Dinge geachtet werden.

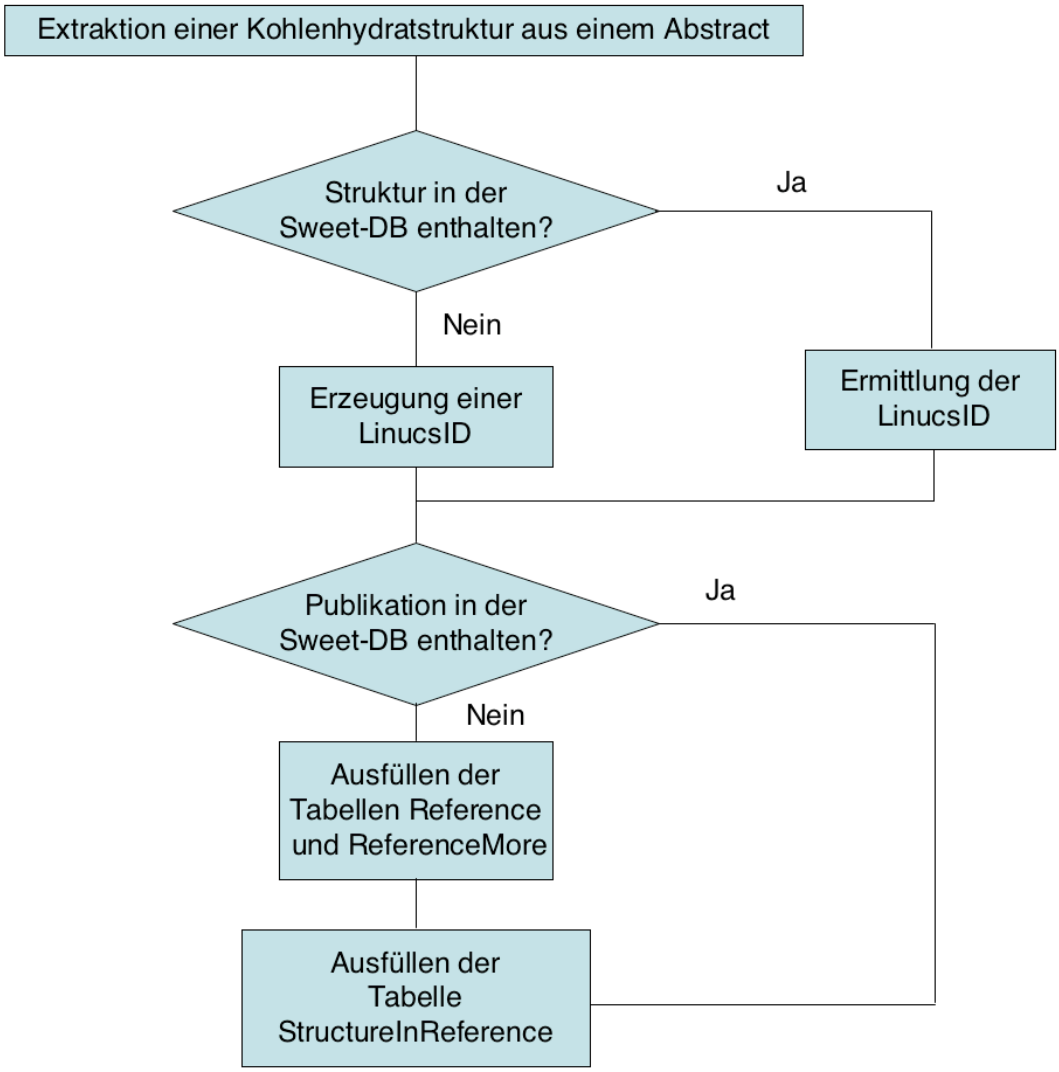


Abbildung 72: Flussdiagramm zum Eintrag einer in einer Publikation gefundenen Strukur

Als erstes muss abgefragt werden, ob eine Struktur schon in der Datenbank gespeichert ist. Wenn nicht, wird als ein eindeutiger Bezeichner die LinucsId erzeugt, der diese Struktur in der Datenbank eindeutig identifiziert. Anschließend wird überprüft, ob die Publikation schon in der Datenbank vorhanden ist. Wenn nicht, erfolgt der Neueintrag in die Tabellen Reference und ReferenceMore. In einem

letzten Schritt wird nun ein Eintrag in die Tabelle StructuresInReference erzeugt, der eine Relation zwischen Publikationen und Strukturen herstellt.

5.6.2 Umsetzung

Ausgehend von den obigen Anforderungen wurden Funktionen erstellt, die aus einem Programm sehr leicht aufgerufen werden können und die entsprechenden Aufgaben erledigen.

Funktionsbezeichnung	Aufgabe
WriteReference	Die Funktion trägt die Daten einer Publikation in die drei Tabellen ein und überprüft ob schon Tabelleneinträge existieren
DeleteRefence	Die Funktion löscht alle Tabelleneinträge zu einer Publikation
writeStructureInReference	Diese Funktion ermöglicht es, einen Eintrag in die Tabelle zur Verwaltung der Strukturen in den einzelnen Referenzen zu schreiben

Tabelle 14: Funktionen der Literatur-Schnittstelle

5.6.3 Benutzung der Routinen über ein Webinterface

Ein erster Test der PHP-Funktionen erfolgt durch ein Webinterface, dass es gestattet die Publikationen der *SWEET-DB* auf einfache Art und Weise zu verwalten. Die PHP-Funktionen finden unter anderem Anwendung bei den Webinterfaces für die Eingabe der NMR (siehe Abschnitt 5.6.6) und Massenspektren (siehe Abschnitt 5.6.7). Das Webinterface kann unter der folgenden URL betrachtet werden: http://www.dkfz-heidelberg.de/spec/sweetdb2/admin/.

5.6.4 Verwendung der C-Schnittstelle

Die Benutzung der C-Schnittstelle erfolgte über das Programm *AUTOREFERENCE*, dessen Beschreibung im Abschnitt 5.7.4 erfolgt.

5.6.5 Ergebnis

Mit Hilfe der erstellten Funktionen ist eine sehr leichte Methode zur Verfügung gestellt, die es gestattet, auf einfache Art und Weise die Publikationen in der *SWEET-DB* zu verwalten. Es werden alle nötigen Funktionen zum Eintragen, Ändern und Löschen der Tabelleneinträge zur Verfügung gestellt. Die Schaffung der Schnittstellen in C und PHP ermöglicht es, die Verwaltung über ein Webinterface und auch einer Applikation zu realisieren. Im nächsten Kapitel dieser Arbeit wird noch genauer beschrieben wie effizient diese beiden Schnittstellen genutzt werden.

5.6.6 Entwicklung einer dezentralen Arbeitsumgebung zur Verwaltung von Messwerten

Natürlich sind die oben beschriebenen theoretischen Massenspektren kein Ersatz für echt gemessene Daten. Leider gibt es zurzeit keine kostenlos über das Internet verfügbare Datenbank mit gemessenen Spektren, die in die *SWEET-DB* eingelesen werden könnte. Da für eine korrekte Identifizierung der Strukturen der Zugriff auf

Vergleichsspektren essentiell ist, führte dieses zu der Überlegung, Spektren an der Stelle zu sammeln, wo sie gemessen werden. Gleichzeitig sollte dem Forscher eine einfache Möglichkeit geboten werden, seine Messergebnisse relativ einfach zu verwalten. Wenn dieses Massenspektrum direkt in eine lokale Datenbank eingespielt wird, kann das Spektrum nach einer Veröffentlichung durch den Forscher über geeignete Schnittstellen in weitere Datenbanken übernommen werden. Eine gleiche Strategie sollte auch für NMR-Spektren erfolgen, die dann ebenfalls den Datenbestand der *SWEET-DB* erweitern sollen.

5.6.6.1 Entwicklung einer webbasierten Umgebung zur Eingabe von NMR-Spektren

Bei der strukturellen Aufklärung von Oligosacchariden spielt die Aufnahme von NMR-Spektren eine sehr große Rolle, da es die einzige physikalische Messmethode ist, die Konformation der einzelnen Hydroxyl-Gruppen genau zu bestimmen. Im Gegensatz zur Massenspektrometrie ist es hier also möglich, die genaue Sequenz und Art der glykosidischen Bindungen zu ermitteln. In der *SWEET-DB* befinden sich zurzeit 1743 Spektren von 1527 verschiedenen Kohlenhydraten, von denen 491 ^{13}C-Spektren und 1255 ^{1}H-Spektren sind. Da zurzeit 22337 verschiedene Strukturen in der *SWEET-DB* gespeichert sind, bedeutet dies, dass nur für 6,8% der Einträge ein oder mehr Spektren gespeichert sind. Es sollte nun eine Strategie entwickelt werden, diesen Anteil zu erhöhen, damit die Wahrscheinlichkeit größer wird, ein gesuchtes Vergleichsspektrum zu finden. Die Extraktion der Daten geschieht im Augenblick durch zwei Studenten, die manuell Strukturen eingeben und die entsprechenden chemischen chemischen Verschiebungen den jeweiligen Atomen zuordnen. Dazu wird eine Publikation, die sich mit der Strukturaufklärung eines Kohlenhydrates beschäftigt, ausgedruckt und die darin enthaltenen Informationen, wie Spezies oder die ermittelten NMR-Daten, werden extrahiert und in die *SWEET-DB* eingetragen. Da dieses Vorgehen sehr arbeits- und zeitaufwendig und damit sehr teuer ist, sollte untersucht werden, wie eine dezentrale Arbeitsumgebung geschaffen werden kann, die lokal für den Benutzer einen Vorteil bei seiner täglichen Arbeit bringt, aber gleichzeitig auch eine schnelle Veröffentlichung in einer Datenbank gestattet.

5.6.6.2 Anforderungen

Bei der täglichen Routinearbeit im Labor wird eine Reihe von NMR-Messungen vorgenommen und anschließend auch ausgewertet, oder es müssen Vergleichs-spektren für eine direkte Identifikation von Glykanen oder Lipopolysacchariden aus Organismen jederzeit zur Verfügung stehen. Es bietet sich an, die lokal aufgenommenen Spektren vor ihrer Publikation in einer nur dem jeweiligen Benutzer zugänglichen Datenbank zu speichern. Die Datensicherheit und Kontrolle des Eingebenden hatte dabei höchste Priorität und es sollte von vornherein eine Lösung entwickelt werden, die es dem Benutzer gestattet festzulegen, wann die Daten in die *SWEET-DB* übertragen werden. Das Verwalten der Daten sollte auf dem Webserver unserer Abteilung erfolgen. Falls ein Benutzer besonderen Wert darauf legt, soll aber auch eine Installation der Komponenten auf einem lokalen Rechner möglich sein. Einen hohen Stellenwert sollte die automatische Generierung des Eingabeformulars haben, da dies der zeitaufwendigste Schritt der Eingabe ist. Da sich mit Änderung der Topologie des Kohlenhydrates auch die Anzahl und die Bezeichnung der Residuen im Eingabeformular ändert, musste das Eingabeformular mit jeder Änderung der Strukturformel neu aufgebaut werden. Dieses ist sehr wichtig für eine

effiziente Eingabe der Messwerte, da so jedes Atom eindeutig identifiziert werden kann, um einfach die korrespondierenden Shift-Werte eingeben zu können.
Dabei sollte auf Standardkomponenten im Webbereich zurückgegriffen werden. Es sollte nun im Folgenden untersucht werden, ob es möglich ist, mit diesen Mitteln eine dezentrale Speichermöglichkeit von NMR-Spektren entwickeln zu können, deren Datenbestand jederzeit mit den in der *SWEET-DB* gespeicherten Spektren abgeglichen werden kann.

5.6.6.3 Konzeptionelle Anforderungen an Software-Komponenten

Eine wichtige Anforderung bei der Umsetzung dieses Projektes bestand darin, dass keine lizenzrechtlichen oder sonstige Kosten als Probleme auf den Benutzer zukommen, was die Akzeptanz und Verbreitung der Anwendung abträglich wäre. Es wurde daher entweder für den akademischen Bereich freie Software verwendet, oder Software, die als Open Source oder unter der Gnu Public License frei weitergegeben werden kann. Die Wahl fiel dabei auf ein System das aus

- MySQL als Datenbankanwendung
- Apache als Webserver
- PHP als Hypertextpreprocessor

besteht. Falls der Benutzer eine Installation der Komponenten vor Ort wünscht, sind diese drei Software-Pakete auch für ungeübte Benutzer relativ schnell zu installieren, da es hierfür eine Anzahl von Installern gibt. Es wurde allerdings bei der Umsetzung wert darauf gelegt, dass die Standardkonfiguration dieser Programme nicht geändert werden musste, und es sollten auch keine zusätzlichen Softwarepakete installiert werden müssen.
Die Arbeitsumgebung sollte auch keine zu großen Anforderungen an die Hardware stellen. So reicht ein Computer, der über einen Pentium-Prozessor, 64 MByte Speicher und einer Festplatte ab einem GByte verfügt, als Server für die Arbeitsumgebung aus.

5.6.6.4 Schutz der eingegeben Daten und Spektren

Da es sich bei der Eingabe der Kohlenhydrate um aktuelle Forschung handelt, und Ergebnisse und Messwerte nicht vor einer Publikation der Öffentlichkeit zugänglich gemacht werden können, sollte eine strikte Zuordnung der eingegebenen Datensätze durch Eingabe eines Benutzernamens und einem dazugehörenden Passwort erfolgen.

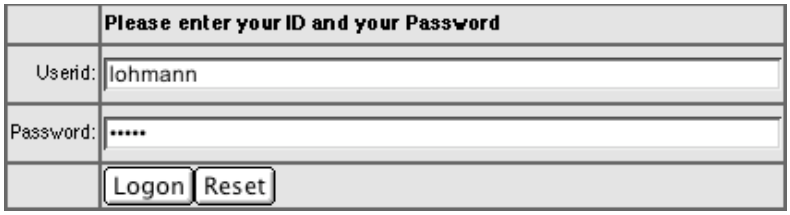


Abbildung 73: Logon in die Arbeitsumgebung

Die Benutzernamen und Passwörter können beliebig lang sein, und es wird zwischen Groß- und Kleinschreibung unterschieden. Somit ist ein ausreichender Schutz gegenüber dem Ausspähen der Daten von Dritten gegeben. Dieses erfordert

natürlich auch die Möglichkeit das Passwort durch den Benutzer verändern zu lassen. Es erfolgt über den Eintrag <Change Password> im Menü auf der linken Seite. Dort kann über ein Webformular nach Eingabe des alten Passwortes und nach zweimaliger Eingabe des neuen Passwortes eine Änderung durchgeführt werden. Danach sind die Eingaben auch vor dem Systemverwalter gesichert, der die Benutzer anlegt und dabei auch die Passwörter vergibt. Um zu gewährleisten, dass in der Abwesenheit des Benutzers die Daten eingesehen werden, kann sich der Benutzer mittels des Menüpunktes <logout> ausloggen. Aus Sicherheitsgründen wird der Benutzer nach einer Stunde der Inaktivität automatisch ausgeloggt.

Das Eintragen der Daten in das Formular erfolgt in einem zweistufigem Verfahren: Zuerst werden die Daten eingegeben, die relativ allgemein sind. Damit das Formular nicht zu unübersichtlich wird und damit nicht zu viele Informationen auf den Benutzer einströmen, wurde das Formular in größere Abschnitte unterteilt. Im Moment sind drei verschiedene Abschnitte definiert: Je ein Abschnitt für

1. Publikationsbezogene Daten
2. Biologisch relevante Daten
3. Experimentelle Daten und Spektren

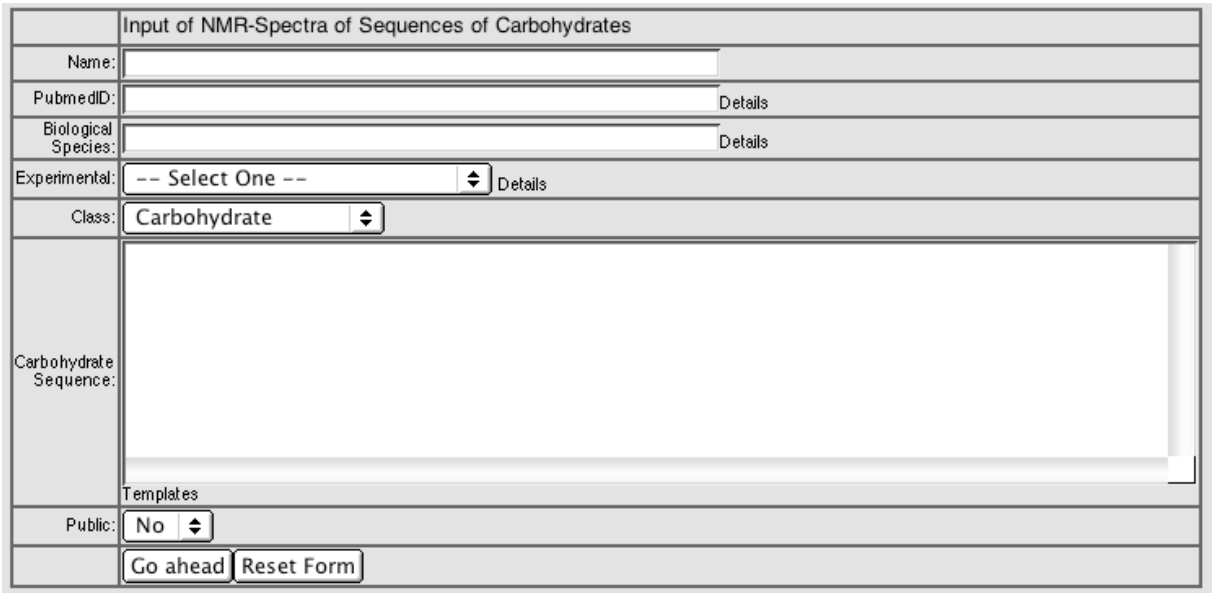


Abbildung 74: Das Webinterface zur Eingabe von NMR-Spektren

Jeder dieser Abschnitte zeigt in der Kurzform des Formulars nur das wichtigste Merkmal der Daten an:

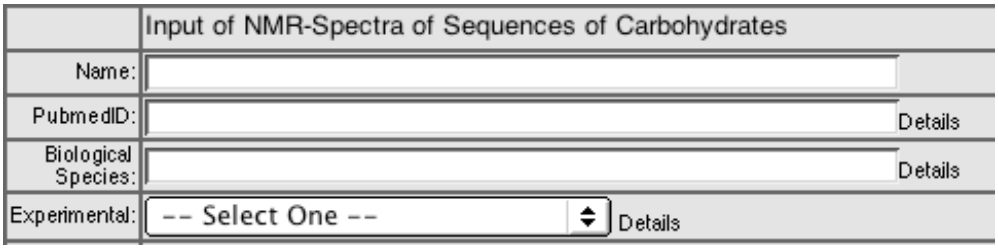


Abbildung 75: Eingabemaske mit den übergeordneten Themen

Durch Anklicken des <Details>-Buttons, der sich am Ende der jeweiligen Zeile befindet, kann der Benutzer sich alle Daten des Abschnittes anzeigen lassen. Als erstes sind hier die publikationsbezogenen Daten des Eintrags zu nennen. Nachdem die Daten und Messwerte publiziert worden sind, können hier die folgenden Daten eingetragen werden:

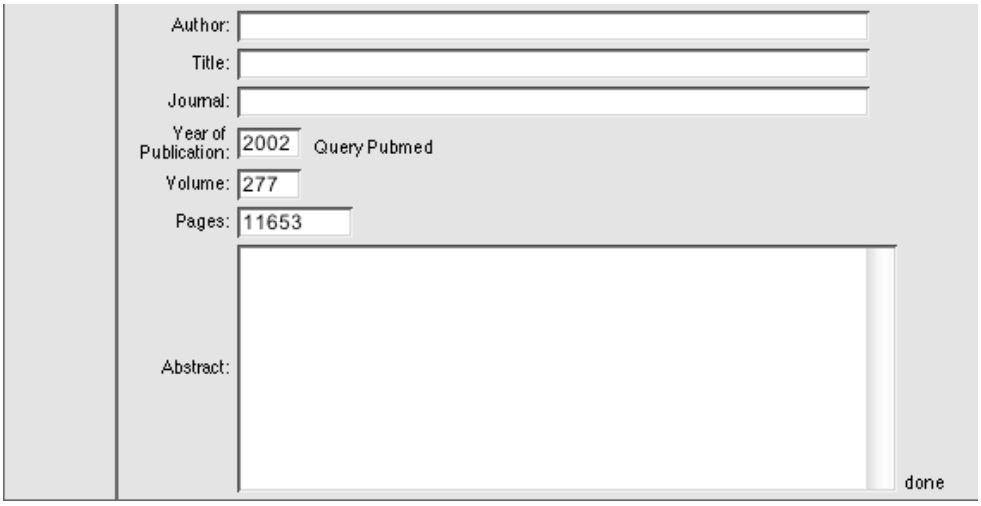


Abbildung 76: Webformular zur Eingabe der bibliographischen Daten

Dieser Vorgang muss für den Benutzer so einfach und effizient wie möglich gestaltet werden. Es sollten bei der Eingabe keine Tippfehler, Zahlendreher oder Ähnliches passieren, da sich diese Fehler bis in den Datenbestand der *SWEET-DB* schleppen würden, und nur mit großem Aufwand wieder korrigiert werden könnte. Es besteht daher die Möglichkeit eine Suche in den Daten der *PUBMED* durchzuführen, um die entsprechende Publikation auszuwählen. Dazu muss nur eine beliebige Kombination der folgenden Werte eingegeben werden:

1. Autor
2. Erscheinungsjahr
3. Ausgabe
4. erste Seite des Artikel

Nach Drücken des <Query Pubmed>-Buttons wird eine Anfrage an die Datenbank der *PUBMED* generiert und die Ergebnisse werden in einem neuen Fenster dargestellt.

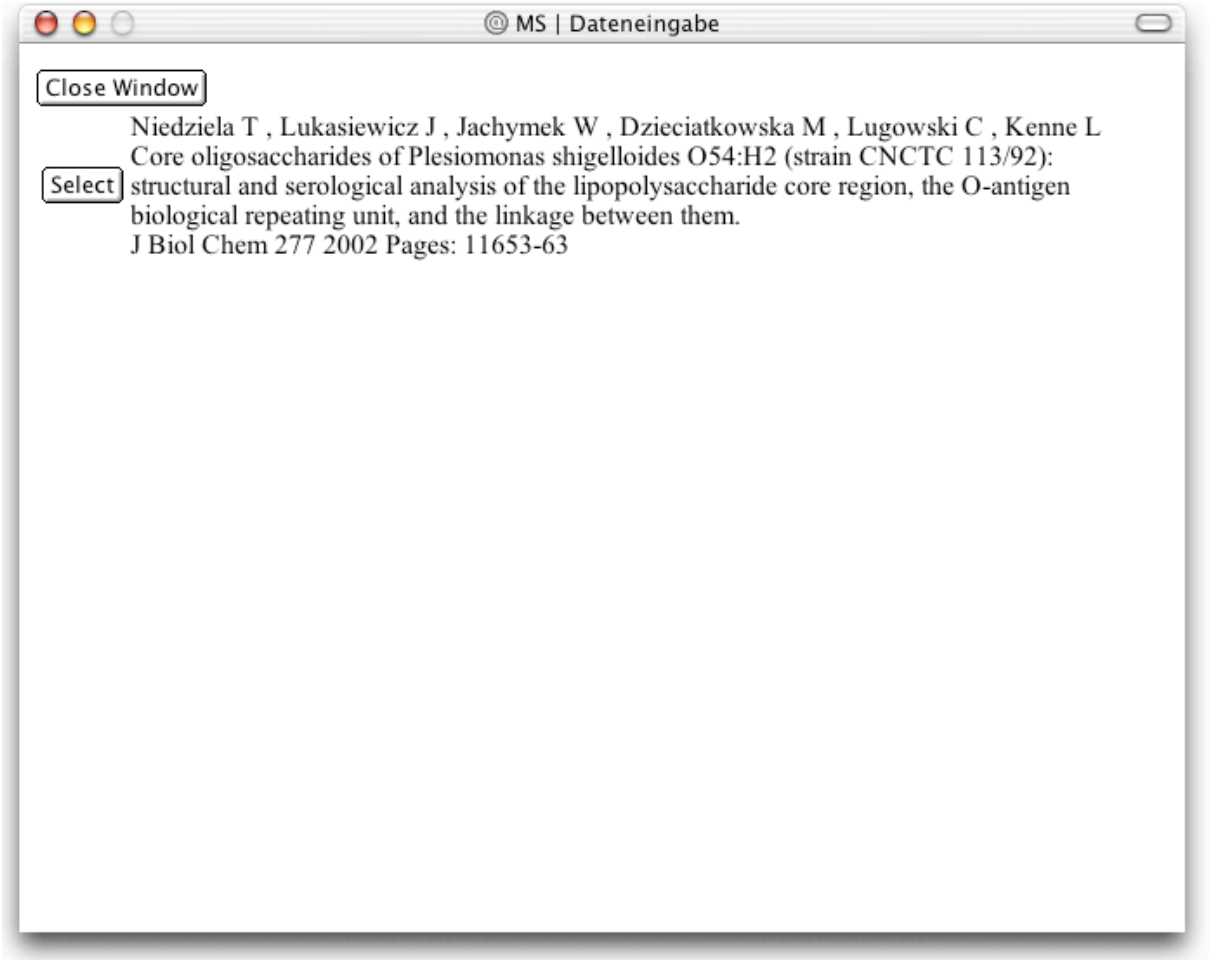


Abbildung 77: Ergebnis der *PUBMED*-Anfrage

Aus diesen Daten kann der Benutzer seinen Artikel auswählen, und die Daten werden sofort in die jeweiligen Felder der Eingabemaske übernommen.

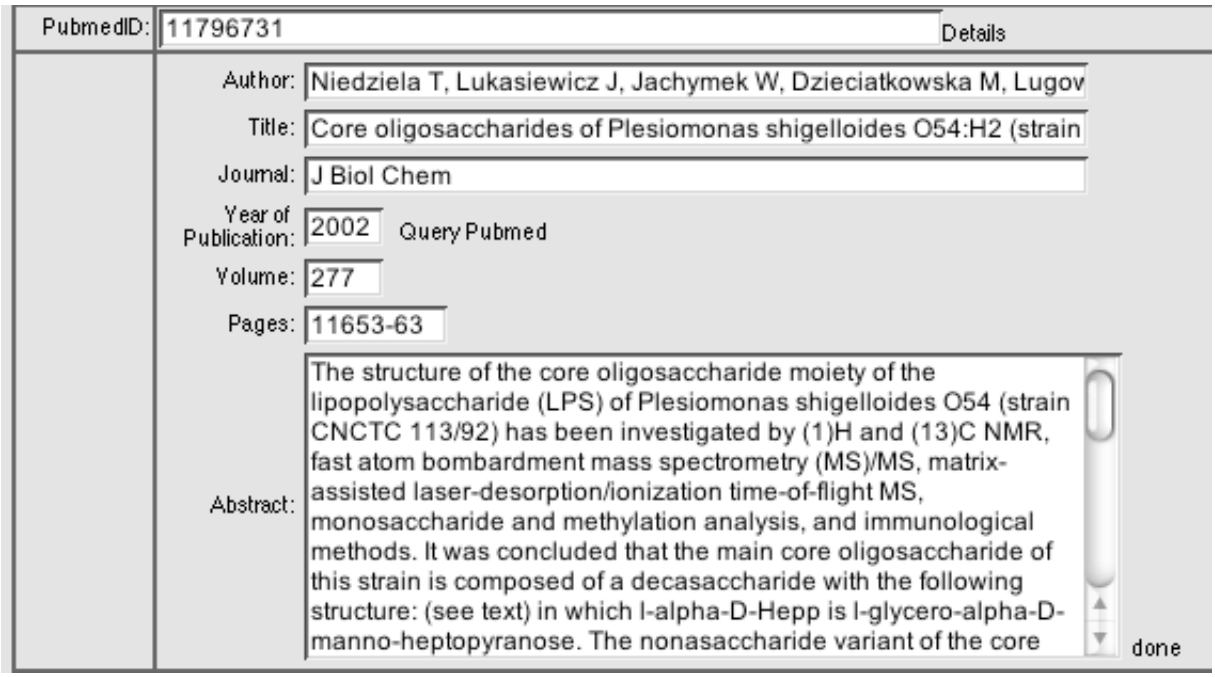


Abbildung 78: Webformular zur Eingabe der bibliographischen Daten

Durch Klicken des <done>-Button kann dieser Vorgang beendet werden, und die Eingabe kann fortgesetzt werden. Nachdem auf diese Art und Weise alle Daten eingegeben worden sind, kann durch Klicken des <go ahead>-Button versucht werden das Formular zur Eingabe der Messwerte zu erstellen. Dabei werden die bis dahin eingegeben Daten einer Prüfung unterzogen. Sollten Daten fehlen oder fehlerhaft sein, wird dieses dem Benutzer im oberen Bereich der Eingabemaske angezeigt.

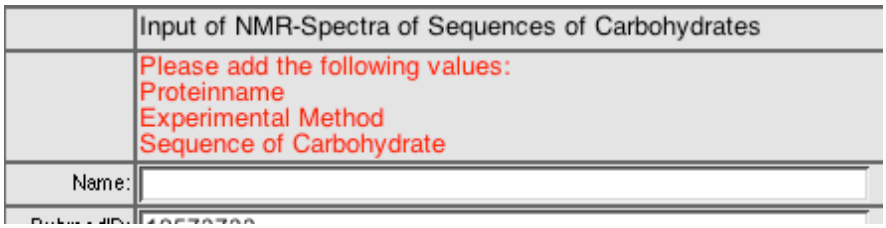


Abbildung 79: Fehlermeldungen nach Überprüfung der Eingabe

5.6.6.5 Strukturangepasste Eingabe der Messwerte

Die Struktureingabe erfolgte wie bei allen Programmen, die in unserem Arbeitskreis entstanden sind, durch die erweiterte IUPAC-Nomenklatur, wie sie in der Einleitung beschrieben ist.

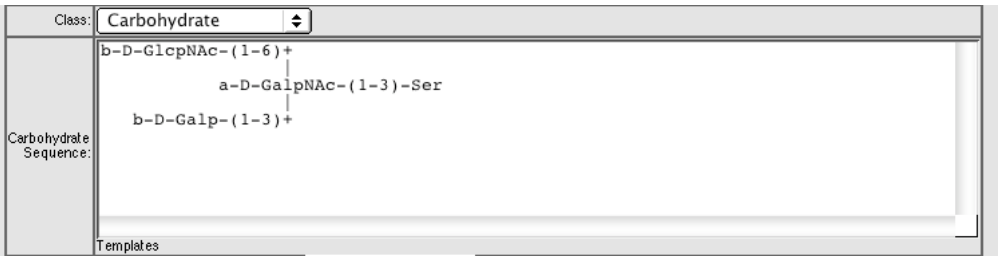


Abbildung 80: Eingabe einer Struktur in der erweiterten IUPAC-Nomenklatur

Daraus wird durch das CGI-Skript *MAKENMRFORM* in Abhängigkeit von der gewählten Spektrenart und von der eingegebenen Struktur ein Formular generiert, in dem alle Atome der jeweiligen Residuen aufgelistet werden. Dazu wird durch das Programm *SWEET-II* eine Datei mit allen enthaltenen Atomen erzeugt, und anschließend werden den Atomen mit dem Programm *UMF-KONVERTER,* das von Martin Frank und Thomas Lütteke in unserer Abteilung erstellt worden ist, die korrekten Atombezeichnungen wie C1 oder H6-ax zugeordnet. Im Moment besteht die Möglichkeit, einen Shift-Wert und eine Kopplungskonstante zu jedem Atom anzugeben. Außerdem war es nötig, die Bezeichnungen für die Atome auch nachträglich noch anpassen zu können. So ist es dem Benutzer möglich, die Bezeichnungen für die Atome zu vergeben, wie sie in der Publikation angegeben sind.

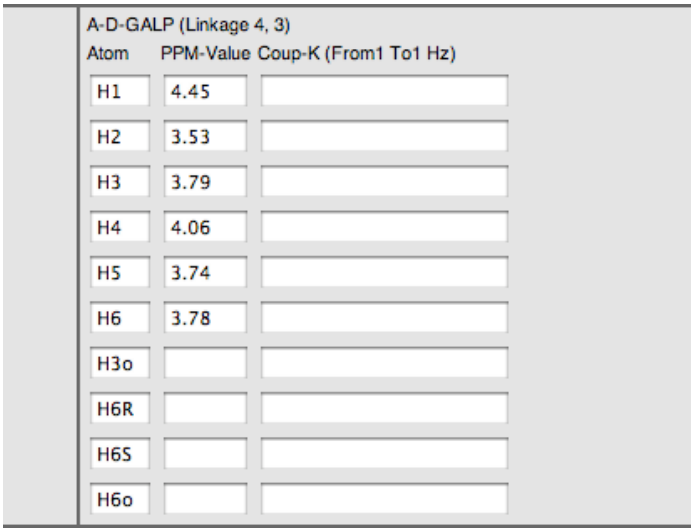


Abbildung 81: Formular zur Eingabe der NMR-Shifts

So werden in den Publikationen die Bezeichnungen für einige Wasserstoff-Atome sehr unterschiedlich gewählt. In einigen Publikationen wird z.B. das H-6R und H-6S auch einfach H-6 und H-6'[114, 115] genannt, in anderen wird es mit H-6a und H-6b[116] bezeichnet. Die von dem Programm *UMF-KONVERTER* vergebenen Bezeichnungen können daher nachträglich einfach verändert werden, und werden zusammen mit den Shift-Werten in die MySQL-Datenbank geschrieben, so dass sie bei einer Übertragung in die *SWEET-DB* zur Verfügung stehen.
Sind alle vorhandenen Messwerte angegeben worden, kann die Eingabe durch Klicken des <save data>-Buttons abgeschlossen werden. Diese Daten werden dann lokal in die MySQL-Datenbank geschrieben und nach Klicken des <list entry>-Button in einer Liste angezeigt.

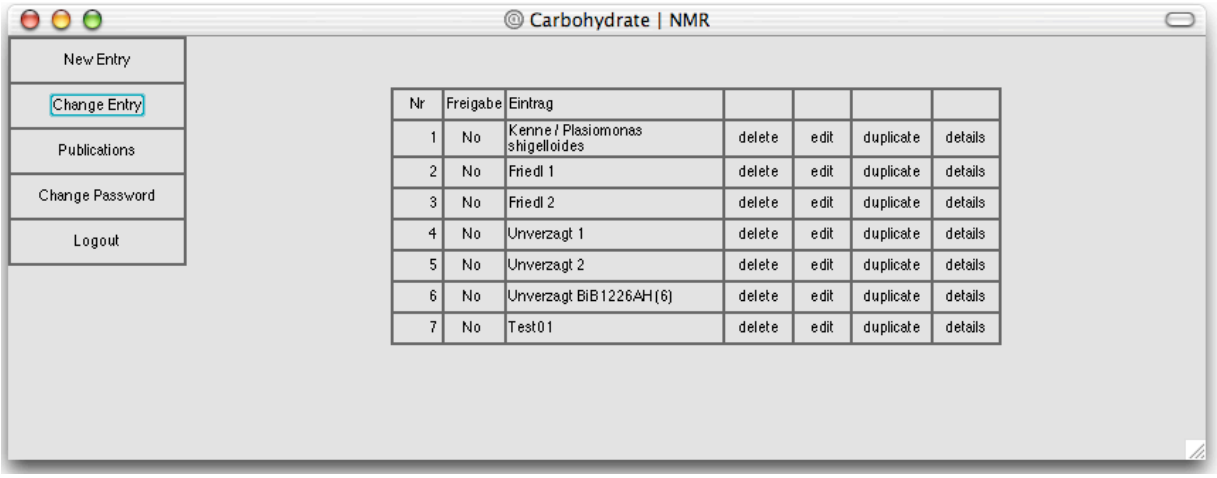


Abbildung 82: Listendarstellung der eingetragenen Spektren

Ausgehend von dieser Liste kann der Benutzer die eingegebenen Daten visualisieren. Er kann auch die Messwerte und Daten nachträglich editieren oder für eine Neueingabe duplizieren. Gerade der Menüpunkt <duplizieren> ist sehr wichtig, da sehr häufig Strukturen eingeben werden müssen, die sich nur in einzelnen Residuen unterscheiden. Die identischen Daten wie Spezies und die publikationsbezogenen Daten werden aber kopiert und müssen nicht erst aufwendig wieder neu eingegeben werden.

5.6.6.6 Editierung der Messwerte

Natürlich kann es immer wieder zu einer nachträglichen Änderung der Daten kommen. Nachdem der Benutzer den <edit>-Button in der obigen Liste angeklickt hat, wird dasselbe Formular angezeigt, wie es auch zur Neueingabe der Daten verwendet wird. Allerdings sind hier die Felder schon mit den vorher eingegebenen Daten ausgefüllt. Diese kann der Benutzer nun beliebig ändern und diese Änderungen auch anschließend speichern.

5.6.6.7 Visualisierung der Messwerte

Natürlich muss es für den Benutzer auch die Möglichkeit geben, seine eingegebenen Daten anzusehen. Dazu war es nötig ein entsprechendes Visualisierungs-Skript für die Darstellung der eingegebenen Shift-Werte zu entwickeln. Die Aufgabe dieses Skripts bestand darin, eine Grafik dieser Shift-Werte zu erzeugen, die einen einfachen Überblick über das Spektrum gestattet. Da der gemessene Kurvenzug nicht zur Verfügung steht, sollte die Ausgabe als Strich-Spektrum erfolgen.

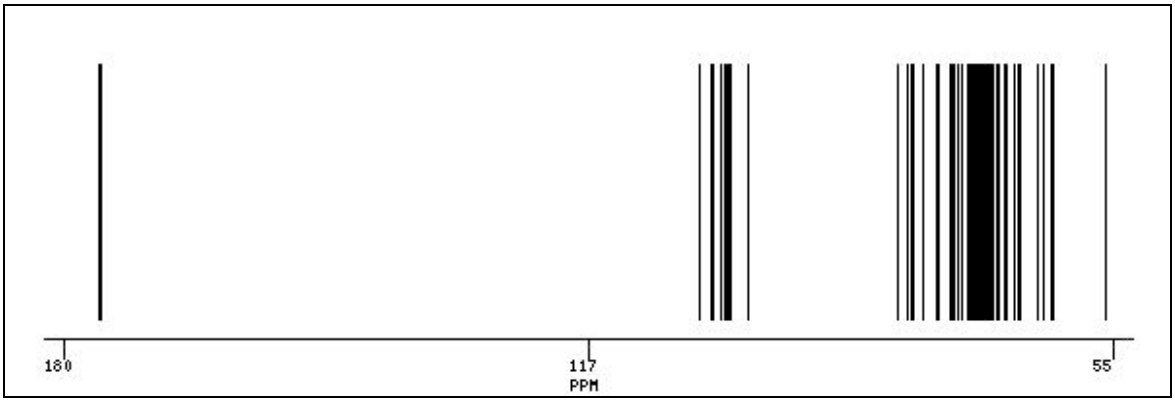
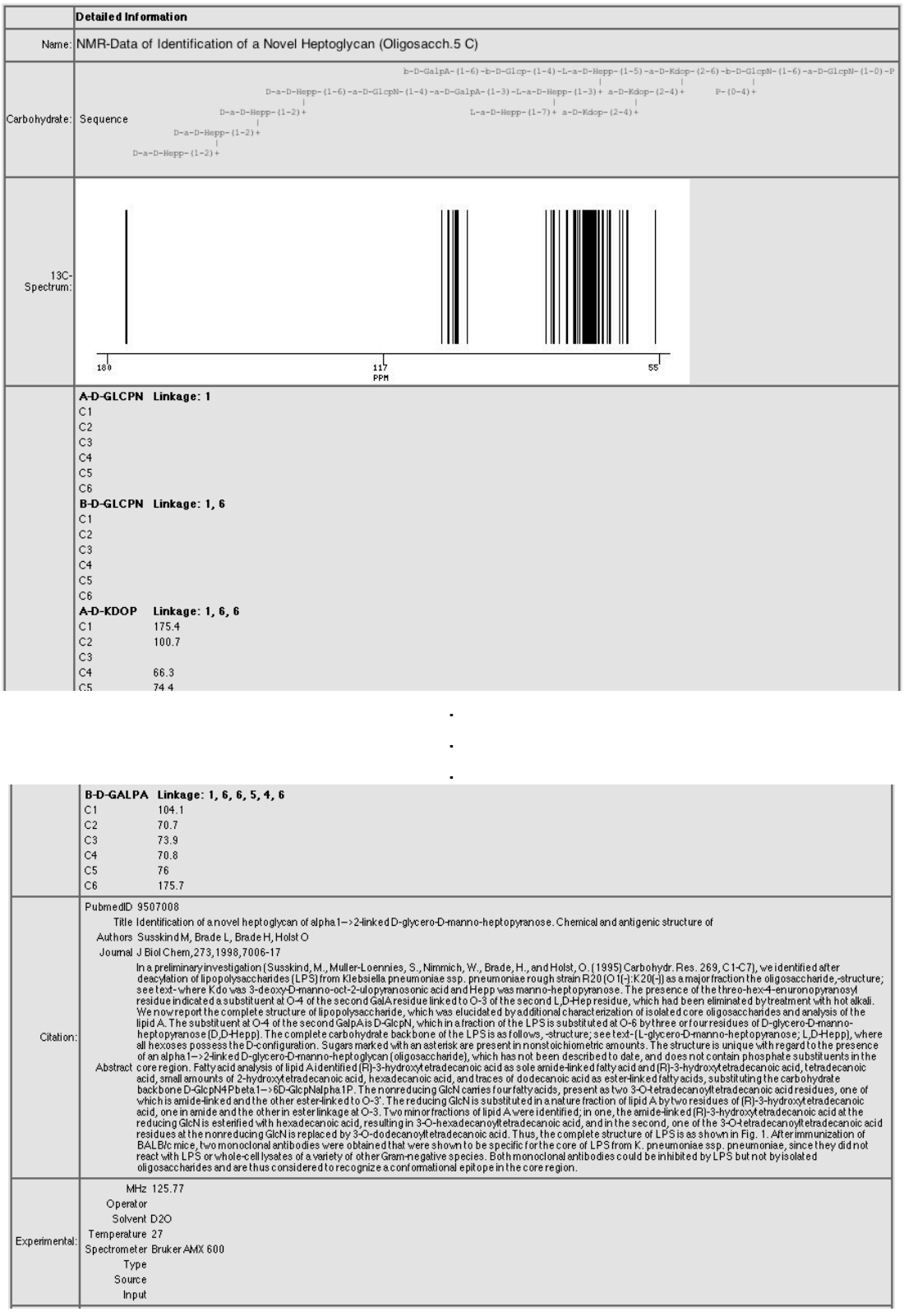


Abbildung 83: Automatisch erzeugte grafische Darstellung der NMR-Daten

Die so erstellten Darstellungen können dann ohne Probleme in eine Gesamtdarstellung des Datensatzes mit einbezogen werden. Man erhält so die folgende Darstellung:



	Detailed Information
Name:	NMR-Data of Identification of a Novel Heptoglycan (Oligosacch.5 C)
Carbohydrate:	Sequence
13C-Spectrum:	180 117 55 PPM

	A-D-GLCPN	**Linkage: 1**
	C1	
	C2	
	C3	
	C4	
	C5	
	C6	
	B-D-GLCPN	**Linkage: 1, 6**
	C1	
	C2	
	C3	
	C4	
	C5	
	C6	
	A-D-KDOP	**Linkage: 1, 6, 6**
	C1	175.4
	C2	100.7
	C3	
	C4	66.3
	C5	74.4

.
.
.

	B-D-GALPA	**Linkage: 1, 6, 6, 5, 4, 6**
	C1	104.1
	C2	70.7
	C3	73.9
	C4	70.8
	C5	76
	C6	175.7

Citation:	PubmedID 9507008 Title Identification of a novel heptoglycan of alpha.1–>2-linked D-glycero-D-manno-heptopyranose. Chemical and antigenic structure of Authors Susskind M, Brade L, Brade H, Holst O Journal J Biol Chem, 273, 1998, 7006-17 Abstract In a preliminary investigation (Susskind, M., Muller-Loennies, S., Nimmich, W., Brade, H., and Holst, O. [1995) Carbohydr. Res. 269, C1-C7], we identified after deacylation of lipopolysaccharides (LPS) from Klebsiella pneumoniae ssp. pneumoniae rough strain R20 [O1(-):K20(-)] as a major fraction the oligosaccharide,-structure; see text- where Kdo was 3-deoxy-D-manno-oct-2-ulopyranosonic acid and Hepp was manno-heptopyranose. The presence of the threo-hex-4-enuronopyranosyl residue indicated a substituent at O-4 of the second GalA residue linked to O-3 of the second L,D-Hep residue, which had been eliminated by treatment with hot alkali. We now report the complete structure of lipopolysaccharide, which was elucidated by additional characterization of isolated core oligosaccharides and analysis of the lipid A. The substituent at O-4 of the second GalpA is D-GlcpN, which in a fraction of the LPS is substituted at O-6 by three or four residues of D-glycero-D-manno-heptopyranose (D,D-Hepp). The complete carbohydrate backbone of the LPS is as follows, -structure; see text-(L-glycero-D-manno-heptopyranose; L,D-Hepp), where all hexoses possess the D-configuration. Sugars marked with an asterisk are present in nonstoichiometric amounts. The structure is unique with regard to the presence of an alpha.1–>2-linked D-glycero-D-manno-heptoglycan (oligosaccharide), which has not been described to date, and does not contain phosphate substituents in the core region. Fatty acid analysis of lipid A identified (R)-3-hydroxytetradecanoic acid as sole amide-linked fatty acid and (R)-3-hydroxytetradecanoic acid, tetradecanoic acid, small amounts of 2-hydroxytetradecanoic acid, hexadecanoic acid, and traces of dodecanoic acid as ester-linked fatty acids, substituting the carbohydrate backbone D-GlcpN4Pbeta.1–>6D-GlcpNalpha1P'. The nonreducing GlcN carries four fatty acids, present as two 3-O-tetradecanoyltetradecanoic acid residues, one of which is amide-linked and the other ester-linked to O-3'. The reducing GlcN is substituted in a nature fraction of lipid A by two residues of (R)-3-hydroxytetradecanoic acid, one in amide and the other in ester linkage at O-3. Two minor fractions of lipid A were identified; in one, the amide-linked (R)-3-hydroxytetradecanoic acid at the reducing GlcN is esterified with hexadecanoic acid, resulting in 3-O-hexadecanoyltetradecanoic acid, and in the second, one of the 3-O-tetradecanoyltetradecanoic acid residues at the nonreducing GlcN is replaced by 3-O-dodecanoyltetradecanoic acid. Thus, the complete structure of LPS is as shown in Fig. 1. After immunization of BALB/c mice, two monoclonal antibodies were obtained that were shown to be specific for the core of LPS from K. pneumoniae ssp. pneumoniae, since they did not react with LPS or whole-cell lysates of a variety of other Gram-negative species. Both monoclonal antibodies could be inhibited by LPS but not by isolated oligosaccharides and are thus considered to recognize a conformational epitope in the core region.
Experimental:	MHz 125.77 Operator Solvent D2O Temperature 27 Spectrometer Bruker AMX 600 Type Source Input

Abbildung 84: Detailierte Anzeige des Datensatzes

Der Benutzer hat so eine einfache Möglichkeit seine Eingaben auf Vollständigkeit zu überprüfen. Er kann auch zum Archivieren den Inhalt des Browserfensters ausdrucken und anschließend abheften. Gleichzeitig besteht aber auch die Möglichkeit die Einträge der Datenbank für eigene Zwecke zu verlinken. So kann z.B. auf der Homepage des Instituts ohne weiteres ein Link angegeben werden, der es gestattet, Daten passend zu einer Publikation anzuzeigen.

5.6.6.8 Datentransfer zur *SWEET-DB*

Der Datentransfer sollte durch bestehende Internet-Protokolle erfolgen, da so eine einfache und effektive Übertragung gewährleistet ist. Die Verwendung des Port 80 (Übertragung durch den Webserver) bietet sich an, da in der Regel auch bei bestehenden Firewalls dieser Port meist freigegeben ist. Andernfalls lässt er sich gefahrlos freigeben und so ist gewährleistet, dass der Port 3306, der für Anfragen an den MySQL-Server benötigt wird, weiterhin hinter der Firewall gesichert ist, und andere laufende Datenbankanwendungen nicht ausgespäht oder sabotiert werden können. Es wurden deshalb PHP-Skripte entwickelt, die es gestatten, die gewünschten Daten in Form eines XML-Containers zu erhalten. Eine schematische Darstellung dieser XML-Container ist im Appendix B dieser Arbeit zu sehen. Durch eine strikte Festlegung auf XML als Grundlage für den Datenaustausch ist gewährleistet, dass auch bei zukünftigen Erweiterungen der Datenstruktur oder bei Wegfall einzelner Bestandteile eine Kompatibilität der einzelnen Versionen bleibt. Es werden jeweils immer nur die vorhandenen oder benötigten Daten aus dem Container extrahiert.
Dieser Vorgang soll natürlich vollautomatisch erfolgen und nur von der Serverseite aus möglich sein. Dabei sind zwei Parameter zu berücksichtigen, die verhindern, dass zum einen Daten ausgespäht werden und zum andern auch verhindern, dass Daten immer wieder doppelt ausgelesen werden. Die Freigabe zur Veröffentlichung erfolgt allein durch den Benutzer. Die Auswahl des Menüpunktes <Ja> im Datenfeld <Freigabe> führt dazu, dass der Datensatz bei nächster Gelegenheit durch die *SWEET-DB* ausgelesen werden kann. Die Häufigkeit dieses Auslesens ist frei einstellbar und erfolgt im Moment einmal wöchentlich. Sind die Daten einmal in die *SWEET-DB* eingelesen, so wird in der lokalen Datenbank ein internes Flag gesetzt, so dass der Datensatz kein zweites Mal ausgelesen wird.

5.6.7 Entwicklung einer webbasierten Umgebung zur Eingabe von Massenspektren

Mit der immer größer werdenden Bedeutung dieser Messmethode für die Aufklärung von Glykanen, Lipopolysacchariden oder Proteinglykosylierungen[54, 117] sollte der Datenbestand der *SWEET-DB* um diese Spektren erweitert werden. Damit besteht eine bessere Möglichkeit, nach Spektren und Massenpeaks suchen zu können. Damit es möglich ist, auch gemessene Spektren in der Datenbank abzulegen, sollte für die Eingabe und Verwaltung dieser Messdaten ebenfalls eine webbasierte Arbeitsumgebung geschaffen werden, die es ermöglicht Spektren und die dazugehörenden Daten dezentral zu speichern und bei Bedarf freizugeben, damit diese in anderen Datenbanken gespeichert werden können.
Dabei wurde von einem typischen Forschungsprojekt der Glykobiologie ausgegangen, das zum Ziel hat, die Glykosylierungen eines Proteins mit Hilfe der Massenspektrometrie zu untersuchen[116, 118]. Bei der Analyse dieser Messmethoden ergab sich eine große Anzahl von Parametern, die in allen Publikationen enthalten waren. Daher müssen sie auch zusammen mit dem Spektrum gespeichert werden, damit sie einfach wieder zur Verfügung stehen. Dazu gehörten Angaben zur Taxonomie, Gewebeart oder Spezies aus dem das Kohlenhydrat gewonnen worden ist. Des Weiteren sollte abgespeichert werden, welche Art von Protein untersucht wurde und in welcher Publikation das Spektrum zum Schluss veröffentlicht worden ist.

5.6.7.1 Anforderungen

Um einen Schutz der Daten zu gewährleisten, wird die Eingabe einer Benutzerkennung und eines dazugehörenden Passwortes verlangt. Das Passwort kann jederzeit geändert werden, und nachdem der Benutzer eine Stunde lang keine Eingabe mehr gemacht hat, sollte eine automatische Abmeldung erfolgen.
Die Authentifizierung des Benutzers erfolgte auf dieselbe Art und Weise wie bei der Eingabe der NMR-Spektren.

5.6.7.2 Technische Umsetzung

Bei der Entwicklung der Arbeitsumgebung zur Eingabe von Massenspektren konnte auf dieselbe Technik zurückgegriffen werden, die auch zur Eingabe der NMR-Spektren benötigt wurde. Dieses gewährleistet ein einheitliches Aussehen der Eingabe-Formulare und eine identische Installation der benötigten Komponenten.

5.6.7.3 Auswertung der Messwert-Datei

In der Regel besteht bei der Software, die bei den Massenspektrometern mitgeliefert wird, eine Möglichkeit, die Messwerte in der Form einer Massenliste als Datei zu exportieren, so dass hier sämtliche Messwerte in dieser Datei an den Datenbankserver übermittelt werden können.

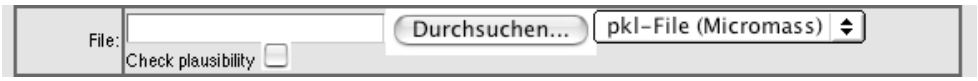


Abbildung 85: Eingabefelder zur Eingabe der Messwertdatei

Natürlich muss eine möglichst große Anzahl von Dateiformaten der einzelnen Hersteller unterstützt werden. Es sind dies zurzeit die Formate der Firmen Micromass (pkl) und Sequest (dta). Außerdem werden zwei neutrale Formate unterstützt. Dabei handelt es sich zum einen um eine einfache Darstellung der Peaks und Intensitäten in reiner ASCII-Form:

```
159.076 77.8
175.084 11.9
175.119 15.8
```

Zum andern wird aber auch das Einlesen von Dateien im CDF-Format unterstützt. Beim CDF-Format handelt es sich um eine Sammlung von Routinen zum Verwalten von Informationen, die sehr gut in Arrays dargestellt werden können. In dieser Dateiform sind auch zusätzliche Parameter, wie die maximale Intensität oder Peakgrenzen, mit abgespeichert. Diese Informationen können zwar mit ausgelesen werden, sie gehen aber bei der anschließenden Speicherung der Daten verloren. Sie sind aber auch für die *SWEET-DB* von keinerlei Interesse.

5.6.7.4 Überprüfung auf Plausibilität

Hat der Benutzer alle Daten eingegeben, so kann er durch Anklicken des <save data>-Buttons die Daten abschicken und sie werden dann in die Datenbank geschrieben. Dabei werden die Daten überprüft und bei Fehlen einzelner Werte erfolgt eine Fehlermeldung.

Abbildung 86: Fehlermeldung der Eingabemaske

Sehr wichtig war dabei die Überprüfung des eingegebenen Spektrums auf Plausibilität. Das heißt es wird überprüft, ob das Spektrum zu der eingegebenen Struktur passt. Dazu wird von dieser Struktur mit Hilfe des Programms *GLYCO-FRAGMENT* ein theoretisches Spektrum berechnet. Dieses wird dann mit dem eingegebenen Spektrum verglichen, wobei mehr als 50 Prozent der Peaks auch darin enthalten sein sollten. Sollte der Wert geringer sein, wird aber nur mit einer Warnmeldung darauf hingewiesen. Da es durchaus sein kann, dass es sich nur um ein schlecht gemessenes Spektrum handelt. Der Benutzer kann diese Meldung dann ignorieren und die Werte werden dann trotzdem in die Datenbank eingetragen.

```
There are only 31 percent of the expected peaks in the
submitted spectra!
```

Dieses Feature kann aber auch von vornherein abgeschaltet werden, wenn feststeht, dass das Spektrum unter Bedingungen aufgenommen worden ist, bei denen nur sehr wenige von den von Domon und Costello beschriebenen Ionen enthalten sind.

5.6.7.5 Visualisierung der Messwerte

Natürlich muss es auch eine detaillierte grafische Darstellung der eingegebenen Daten geben. Dazu ist es nötig, aus dem Spektrum eine Darstellung des Massenspektrums zu berechnen. Dieses führt zu der folgenden Darstellung:

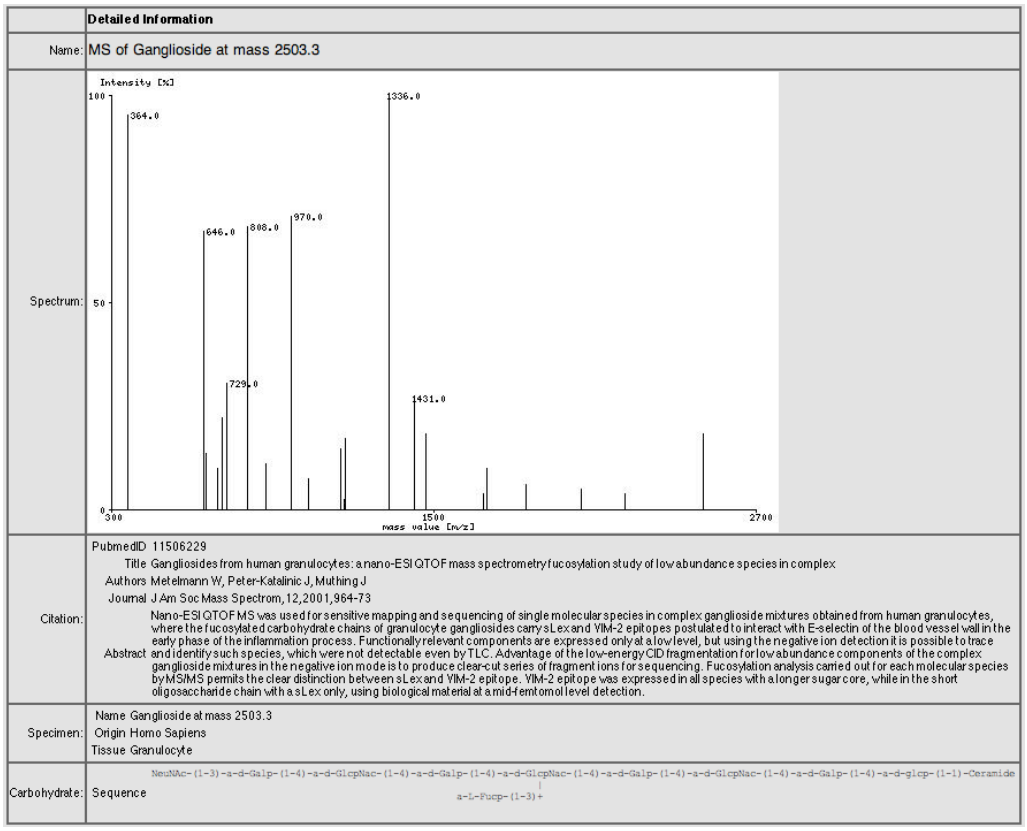


Abbildung 87: Detaillierte Darstellung der einzelnen Daten mit Bild des eingegebenen Spektrums

5.6.7.6 Diskussion und Ausblick

Ein erster Test durch Studenten der Biologie zeigte, dass es mit Hilfe dieser Arbeitsumgebung möglich ist, auch eine größere Anzahl von Spektren einzugeben und zu verwalten. Durch die Verwendung von XML-Containern und etablierten Protokollen zur Übertragung der Daten in die *SWEET-DB* konnte die Anzahl der gespeicherten Spektren in der Zeit von sechs Monaten um 347 Spektren erweitert werden, was einer Steigerung um 15% entspricht. Durch das Webinterface kann der Benutzer seine Daten direkt über das Internet verwalten, und er ist so unabhängig von einem bestimmten Rechner oder Arbeitsplatz. Im Falle einer Publikation kann er seine Daten freigeben, und sie stehen sofort dem Datenbestand der *SWEET-DB* zur Verfügung. Durch die Entwicklung des Formulargenerators können ohne Probleme auch Eingabemasken für ^{31}P- oder ^{17}N-NMR-Spektren zur Verfügung gestellt werden. Um die Eingabe von NMR-Spektren in Zukunft zu beschleunigen, sollte auch die Möglichkeit gegeben sein, die Kurvenzüge der Messwerte als JCAMP-Datei einzulesen. Die Eingabe von NMR- und Massenspektren verkürzt sich durch die Arbeitsumgebung sehr stark, und durch die einfache Übertragungsmöglichkeit an die *SWEET-DB* brauchen Daten nicht doppelt eingegeben werden. Es braucht daher keine aufwendige Kontrolle durch einen Experten erfolgen.
Durch die sorgfältige Entwicklung der Schnittstellen zum Datentransfer ist auch für zukünftige Anwendungen eine Basis geschaffen, die einen weiteren Ausbau des Datenbestandes der *SWEET-DB* gestattet.

5.7 Automatische Annotierung und Klassifizierung von Publikationen

Der Nutzen einer Datenbank steht und fällt mit der Aktualität und der Vollständigkeit der in ihr enthaltenen Daten. In der jüngeren Vergangenheit hat es verstärkt Bemühungen gegeben, die es ermöglichen sollen, einen Text automatisch einem Themengebiet zu zuordnen[119] oder automatisch Informationen aus den Publikationen zu extrahieren[120, 121]. Dabei ist es wichtig, dass die Vollständigkeit der extrahierten Daten ein hohes Niveau hat und es nicht zu einer zu großen Anzahl von falsch positiven Einträgen kommt. Es ist sehr unbefriedigend, bei der Recherche in z.B. der *SWEET-DB* nicht relevante Publikationen angezeigt zu bekommen. Die Anzahl der jährlich veröffentlichten Publikationen hat sich in den letzten dreißig Jahren mehr als verdoppelt. Diese Menge kann nicht mehr manuell verarbeitet werden.

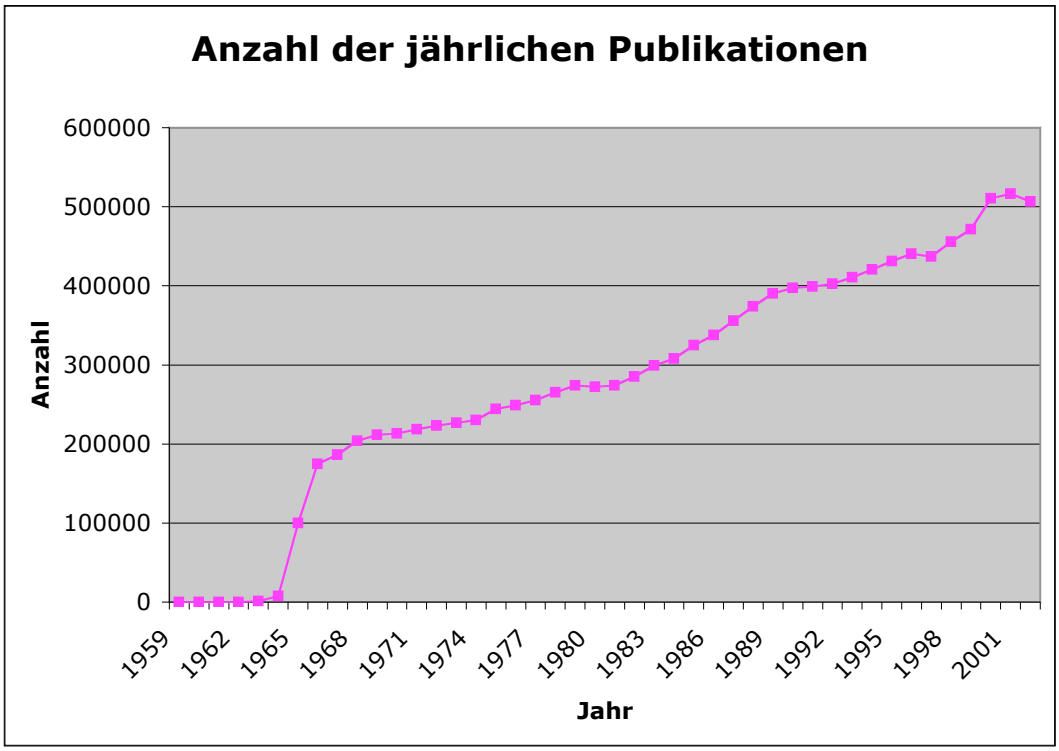


Abbildung 88: Anstieg der jährlichen Publikationen (Quelle: *PUBMED*)

Im Folgenden sollte untersucht werden, inwieweit es möglich ist, diesen Vorgang zu automatisieren. Es sollten verschiedene Strategien entwickelt werden, die es gestatten, die Neuaufnahme der Publikationen in der *SWEET-DB* sehr einfach zu verwalten und neue Einträge thematisch richtig einzuordnen. Um den jetzigen Datenbestand der *SWEET-DB* nicht zu gefährden, sollten die Daten allerdings erst nach Sichtung durch einen Experten in die Datenbank geschrieben werden.

5.7.1 Grundsätzliches Vorgehen

Die *PUBMED* ist für die Recherche von Informationen im biologischen Bereich die Datenbank der Wahl, da in ihr praktisch alle relevanten Publikationen enthalten sind. Es sollte nun in einem ersten Schritt Daten aus der *PUBMED* extrahiert werden. Diese Daten werden zuerst in eine temporäre Datenbank, die *AUTO-SWEET-DB,* geschrieben, da so thematisch nicht relevante Einträge ohne weiteres wieder gelöscht werden können. Dazu war es nötig zwei verschiedene Programme zu

entwickeln. Zum einen war es nötig, ein Programm zu entwickeln, das als im Hintergrund laufender Prozess permanent Rohdaten in eine Datenbank schreibt. Mittels eines zweiten Programms werden diese Daten anschließend gesichtet und in die *SWEET-DB* geschrieben:

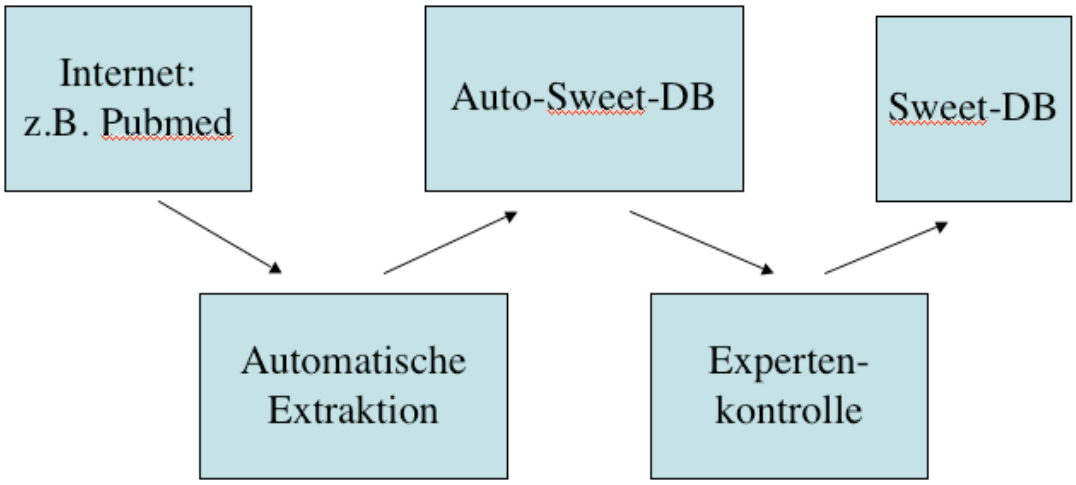


Abbildung 89: Schematische Darstellung des Datenflusses

5.7.2 Trefferquote der gefundenen Daten

Um die Vollständigkeit der gefundenen Daten beurteilen zu können, war es nötig, einen Wert zu definieren, der angibt wie gut, die im folgenden untersuchten Suchstrategien nach denen in der *PUBMED* Publikationen extrahiert worden sind, und auch zu tatsächlichen Einträgen in die *SWEET-DB* führen. Dazu wird der Quotient aus gefundenen Publikationen und in die *SWEET-DB* übernommenen Publikationen gebildet:

$$\text{Trefferquote} = \frac{\text{Publikationen}_{\text{Sweet-DB}}}{\text{Publikationen}_{\text{Auto-Sweet-DB}}}$$

Je größer dieser Quotient ist, desto besser ist die eingesetzte Extraktions-Strategie. Ab einem Wert von 0,99 kann davon ausgegangen werden, dass die Suchstrategie vollautomatisch durchgeführt werden kann. Strategien bei denen der Wert kleiner als 0,5 ist, sollten nicht weiter verfolgt werden, da der Aufwand bei der Sichtung zu groß ist im Verhältnis zu den gewonnen Daten.

5.7.3 Eigene Arbeiten

Ausgehend von diesen Anforderungen und Überlegungen wurden nun mehrere Strategien entwickelt, die es gestatten, den Datenbestand auf einfache Art und Weise zu verwalten.

5.7.4 AUTOREFERENCE

AUTOREFERENCE ist ein Programm, das dafür ausgelegt worden ist, als Prozess im Hintergrund zu laufen und permanent die *PUBMED* darauf zu untersuchen, ob neue Einträge vorhanden sind, die relevant sind für die *SWEET-DB*. Dazu wurden

verschiedene Optionen in das Programm eingebaut, die den verschiedenen Aufgabenstellungen gerecht wurden. So hat das Programm zwei verschiedene Modi:

1. Suche nach Trivialnamen
2. Suche nach Kohlenhydratstrukturen

Mit Hilfe dieser Anwendung kann nun ein relativ großer Bestand der *PUBMED* durchsucht werden. Bei der Suche nach Trivialnamen erfolgt die Suche über die definierte Schnittstelle des *PUBMED*-Webservers[1]. Dieser Schnittstelle wird nun als Parameter der Trivialname und der gewünschte Suchzeitraum übergeben. Als Antwort erhält man einen XML-Container mit einer Liste von Publikationen, die durch die *PUBMED*-ID eindeutig identifiziert werden können. Die Antwort wird ausgewertet, und die gefundenen Publikationen werden eine nach der anderen in die *AUTO-SWEET-DB* geschrieben. Bei der Suche nach Kohlenhydratstrukturen existiert kein vordefiniertes Suchwort. Es wird daher einfach als Suchkriterium die *PUBMED*-ID, die einem fortlaufenden Index in der Datenbank entspricht, übergeben. Diese ID wird sukzessive um eins erhöht, so dass alle Einträge überprüft werden können.

5.7.5 REFERENCE

Nachdem die Daten von der *PUBMED* in die *AUTO-SWEET-DB* geschrieben worden sind, muss es ohne größere Probleme möglich sein, dass die gefundenen Datensätze von einem Experten gesichtet werden können, um einerseits nicht relevante Datensätze zu löschen, und um andererseits die relevanten Datensätze in die *SWEET-DB* zu transferieren. Dieses sollte so effizient wie möglich geschehen, um nicht unnötig Ressourcen zu binden. Dazu wurde die Applikation *REFERENCE* entwickelt. Die Hauptaufgabe dieser Anwendung bestand darin, die Daten, die automatisch ermittelt worden sind, aufzulisten und danach entweder zu löschen oder in die *SWEET-DB* zu schreiben. Dazu wurde die in Abschnitt 5.6.1 beschriebene C-Schnittstelle zum Eintragen von Publikationsdaten in die *SWEET-DB* benutzt.
Das Interface wurde so benutzerfreundlich wie möglich gestaltet. Im oberen Teil ist eine Liste aller automatisch extrahierten Literaturstellen zu sehen. Im unteren ist es möglich sich alle Daten, die zu einer Publikation gehören, sich anzeigen zu lassen:

[1] http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

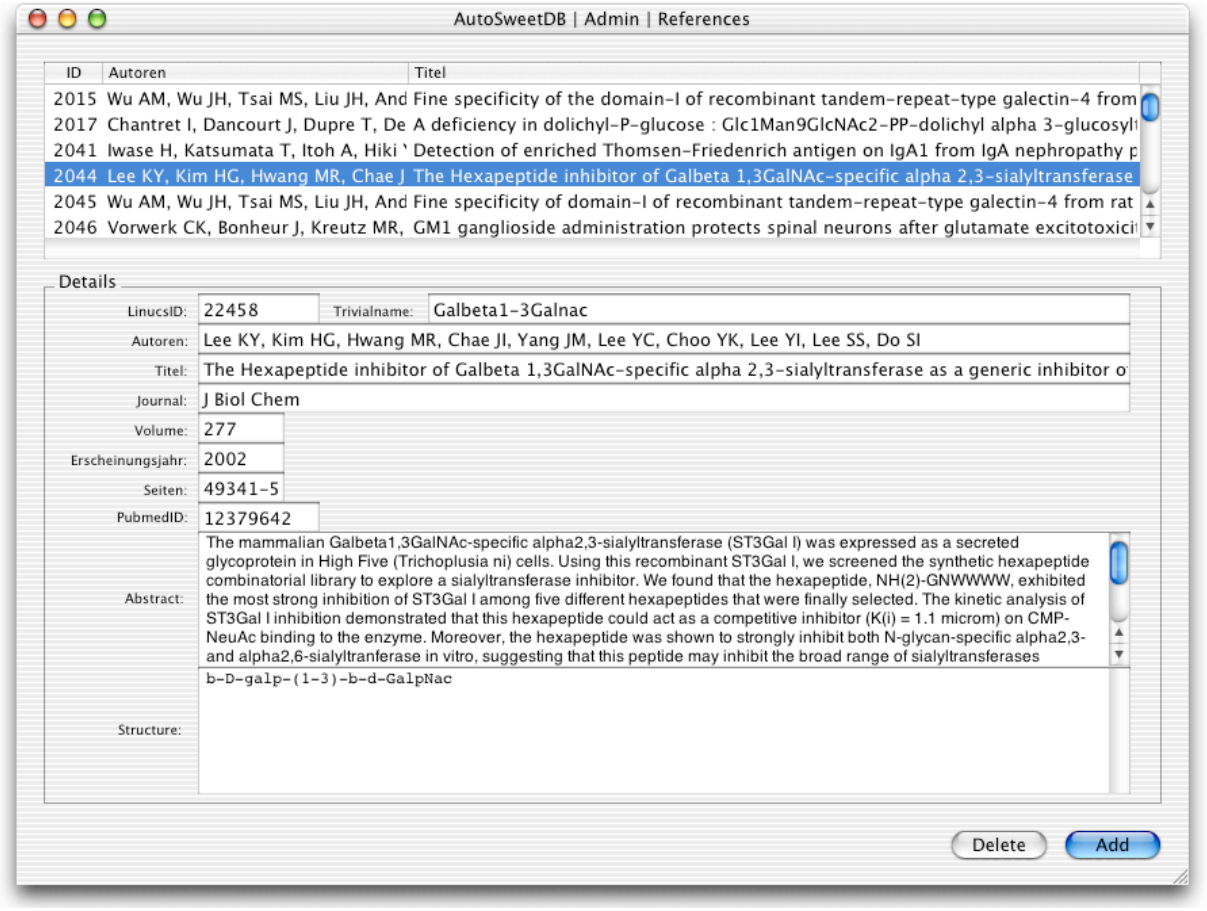


Abbildung 90: Interface des Programms *REFERENCES*

Mit Hilfe der beiden Buttons <Delete> und <Add> kann sehr schnell entschieden werden, ob eine Publikation übernommen werden soll oder nicht. In ersten Fall werden die Daten aus der *AUTO-SWEET-DB* gelöscht und in die *SWEET-DB* geschrieben. Damit stehen sie sofort den Benutzern der *SWEET-DB* zur Verfügung. Im zweiten Fall werden die Daten einfach gelöscht. So kann Publikation für Publikation durchgegangen werden. Der Aufwand für 100 Publikationen liegt im Bereich von 37 Minuten. Wobei in der Regel 88 Prozent der gefundenen Publikationen auch in die *SWEET-DB* eingetragen werden können.

5.7.6 Extraktion von Publikationen aus der *PUBMED*-Datenbank

Die Extraktion der Datensätze von den zu klassifizierenden Publikationen wurde über die von der *PUBMED* zur Verfügung gestellten Retrieval-Schnittstellen gewährleistet. Dazu gibt es zwei verschiedene Schnittstellen bei der Pubmed, die über das Internet erreicht werden können. Mit der Schnittstelle *ESEARCH*[1] besteht die Möglichkeit, Suchanfragen an die *PUBMED* zu stellen. Das Ergebnis wird als Liste von *PUBMED*-Ids geliefert. Übergibt man diese *PUBMED*-ID der Schnittstelle *EFETCH*[2] so erhält man alle Daten, die zu dieser ID in der *PUBMED*-Datenbank gespeichert sind. Der zurück gelieferte Datensatz liegt im XML-Format vor und kann ohne Probleme analysiert werden.
Im vorliegenden Fall wurde auf eine Eigenentwicklung zurückgegriffen, damit es keinerlei Probleme mit Lizenzen oder mit den Statuten der National Institutes of

[1] http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esearch_help.html
[2] http://eutils.ncbi.nlm.nih.gov/entrez/query/static/efetch_help.html

Health (NHI) zur kostenlosen Nutzung der *PUBMED*-Datenbank gibt. Diese Statuten sehen vor, dass es nicht mehr als Fünftausend Zugriffe von einem einzelnen Rechner am Tag geben soll, und der Abstand zwischen zwei Zugriffen soll mehr als drei Sekunden betragen.
Ein weiterer Grund für eine Eigenentwicklung der XML-Routinen war, dass diese Klassen und Methoden in anderen Projekten weiterbenutzt werden können. Die folgenden Tabellen zeigen den schematischen Aufbau der Klassen und die zu Grunde liegenden Methoden, die in den Klassen enthalten sind. Zum einen sind es die Funktionen zur Extraktion des XML-Datensatzes aus dem Datenbestand der *PUBMED*, zum anderen sind es die Funktionen zur Analyse des Datensatzes. Es wurde dabei Wert darauf gelegt, dass sie sehr allgemeingültig für die Analyse auch anderer XML-Datensätze sind. Trotzdem sollte aber auch ein schneller und einfacher Zugriff auf die Elemente des *PUBMED*-Datensatzes gewährleistet sein.

5.7.7 Programmierung der Methoden

Es wurden drei grundlegende Methoden zur Extraktion der Daten aus dem XML-Datensatz benötigt:

Methode	**Beschreibung**
GetXmlEntry	Die Methode liefert den Wert für einen XML-Tag
CountXmlEntry	Die Methode zählt die Einträge für einen XML-Tag
GetXmlNEntry	Die Methode liefert den n-ten Wert für einen XML-Tag

Tabelle 15: Grundlegende Methoden zur Analyse eines XML-Datensatzes

Um den schnellen Zugriff, auf die *PUBMED*-Daten zu erhalten wurden Funktionen implementiert, die aufgerufen, die gewünschten Daten zurückliefern.

Methode	**Beschreibung**
GetAbstract	Die Methode liefert das Abstract der Publikation oder einen String der Länge Null
. . .	
GetYearOfPublicationDate	Die Methode liefert das Jahr, in dem die Publikation veröffentlicht wurde

Tabelle 16: Spezielle Methoden zur Analyse eines Pubmed-Datensatzes

5.7.8 Datenextraktion durch Suche nach Trivialnamen

Auf Grund der strukturbasierten Speicherung der Daten in der *SWEET-DB* ist es nötig, dass eindeutig erkannt wird, zu welcher Struktur eine in der *PUBMED* gefundene Publikation gehört. Dieses bedeutet, dass der eindeutige Identifier, die LinucsID, der in der *SWEET-DB* für jede Struktur vergeben ist, der Publikation eindeutig zugeordnet werden muss. Da in der *PUBMED* keine strukturrelevanten Angaben gespeichert werden, war es nötig, zu erkennen, welche Zuckerstrukturen in dem jeweiligen Artikel behandelt werden.

In einem ersten Versuch wurden einfach alle in der *SWEET-DB* gespeicherten Trivialnamen genommen und mittels des Programms *AUTOREFERENCE* wurde versucht alle Publikationen ab dem Jahr 1960 zu finden, die in einem der *PUBMED* Datenfelder einen dieser Begriffe enthielten. Dabei wurden auch innerhalb kürzester Zeit mehr als Zehntausend Publikationen gefunden und in die *AUTO-SWEET-DB* geschrieben. Bei einer anschließenden Sichtung mit dem Programm *REFERENCE* stellte sich aber sehr schnell heraus, dass mit dieser Methode nur eine relativ kleine Anzahl verwertbarer Artikel gefunden werden konnten. So ergab sich für die Vollständigkeit der Suche:

$$\text{Trefferquote} = \frac{134}{10142} \approx 0{,}01$$

Es mussten also mehr als 99 Prozent der Einträge verworfen werden. Auf der Suche nach den Ursachen stellte sich sehr schnell heraus, dass dieses an der Art der Suchbegriffe lag. Auf der einen Seite führten Begriffe wie Laktose oder Cellulose schnell zu einer großen Anzahl von Artikeln, die nur selten in den Bereich der Glykobiologie einzuordnen sind. Es wurde zum Beispiel jede Publikation aufgeführt, die zu den Themen Laktoseinteloranz[122], Laktose als Hilfsstoff in Tabletten[123] oder Cellulose als Bestandteil von Arzneizubereitungen[124] gehörten. Auf der anderen Seite wurden aber zu sehr eindeutigen Begriffen, wie ‚kolomiktriose stigmastenol'[125] keine Ergebnisse gefunden, da dieser Begriff nur in einer einzigen Publikation verwendet wurde. Es musste also eine neue Suchstrategie überlegt werden.
In einem zweiten Ansatz wurde nun mit einer durch einen Experten editierten Liste in der *PUBMED* gesucht. Die Verwaltung dieser Liste erfolgt durch das Programm *TRIVIALNAMES*. Die Suche wurde wieder mit dem Programm *AUTOREFERENCE* durchgeführt.

```
GQ1b
Gt1c
Gt1b
Gt1a
Gal-Gb4
Gb3 ganglioside
Globo-H
Man8glcnac2
Man7glcnac2
Man4glcnac2
Gb4 glycosphingolipid
Gb3 glycosphingolipid
Gd3 ganglioside
Gd2 ganglioside
Gd1b ganglioside
Gd1a ganglioside
Gm3 ganglioside
Gm2 ganglioside
Gm1 ganglioside
Lewisb
sialyl Lewisa
Lewisa
```

```
Lewisy
glc1Man5GlcNac2
Glc3man9glcnac2
Man3glcnac4
Man6glcnac2
Man5glcnac2
Man3glcnac2
sialyl LewisX
Slex
Lewisx
SDA carbohydrate
glc2Man5GlcNac2
blood group h antigen carbohydrate
blood group a antigen carbohydrate
blood group b antigen carbohydrate
hnk-1 carbohydrate
Galbeta1-3Galnac
```

Mit diesen Begriffen wurden nun alle seit 1999 in der *PUBMED* gespeicherten Publikationen gesucht, die in einem der Datenfelder einen dieser Begriffe enthielten. Die gefundenen Einträge wurden dann in die *AUTO-SWEET-DB* eingetragen und anschließend mit dem Programm *REFERENCE* überprüft und in die *SWEET-DB* überführt. Dabei wurden für den Zeitraum von 1999 bis 2002 insgesamt 944 neue Publikationen in die Datenbank eingetragen. Ursprünglich waren 1062 Publikationen gefunden worden, was bedeutet, dass so eine sehr hohe Güte der Suche von

$$\text{Trefferquote} = \frac{944}{1062} \approx 0{,}88$$

erreicht worden ist. Dieses entspricht einem sehr guten Ergebnis und eine Datenextraktion dieser Art erfolgt nun Monat für Monat. Somit erlaubt dieser Ansatz auf sehr einfache Art und Weise die Publikationen in der *SWEET-DB* auf dem neuestens Stand zu halten. Mittels diesen Ansatzes konnten die Publikationen bis zum Monat Juli 2003 in die *AUTO-SWEET-DB* und nach einer Sichtung in die *SWEET-DB* geschrieben werden. Bei der Durchführung der Suche hat sich herausgestellt, dass eine vollständige Indizierung der einzelnen Journale mit etwa zwei Monaten Verzögerung abgeschlossen ist. Die Durchführung der Suche in der *PUBMED* erfolgt also mit einem Versatz von zwei Monaten.

5.7.9 TRIVIALNAMES

Das Programm *TRIVIALNAMES* erlaubt eine sehr einfache Verwaltung der zu suchenden Trivialnamen mit Zuordnung zu den entsprechenden LinucsID, die in der *SWEET-DB* vergeben ist. So ist gewährleistet, dass zu jeder gefundenen Publikation eine Struktur eindeutig zugeordnet werden kann. Außerdem kann hier auf sehr einfache Weise festgelegt werden, ab wann die Suche in der *PUBMED* erfolgen soll.

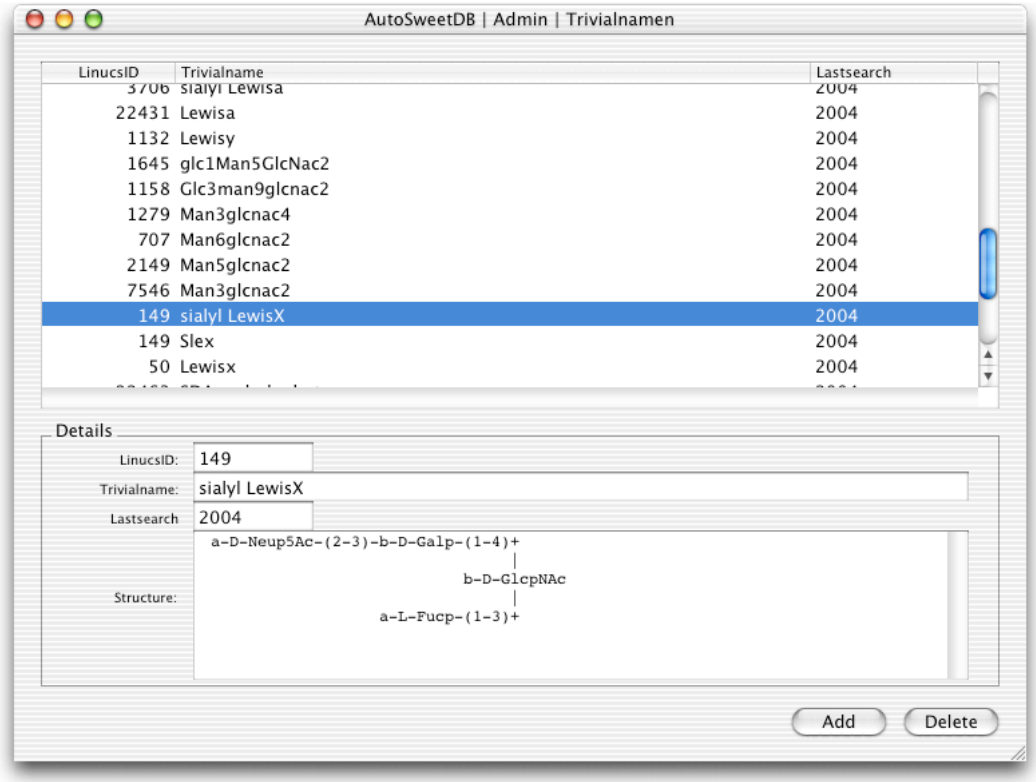


Abbildung 91: Interface des Programms *TRIVIALNAMES*

Sollte ein Eintrag fehlerhaft sein, so kann er ohne Probleme verändert werden.

5.7.10 GETABSTRACTS

In der *SWEET-DB* waren bis zu diesem Zeitpunkt noch keine Abstracts der in der *CARBBANK* und *SUGABASE* untersuchten Publikationen enthalten. Dieses sollte mittels der oben beschriebenen Methoden der Klasse zur Auswertung von XML-Containern realisiert werden. Dazu wurde das Programm *GETABSTRACTS* entwickelt, das nacheinander eine Liste von *PUBMED*-Ids abarbeitet und den zu der *PUBMED*-ID gehörenden Abstracttext in die *SWEET-DB* einträgt. Bei dem Programm *GETABSTRACTS* waren zwei Dinge zu beachten: Zum einen sollte das Programm auch die oben beschriebene Schnittstelle benutzen. Zum anderen waren auch hier wieder die Nutzungsbedingungen der NHI zu beachten. Da die Geschwindigkeit der Datenextraktion keinerlei Rolle spielte, wurde sogar nur alle dreißig Sekunden ein Abstract von der *PUBMED* angefordert und in die *SWEET-DB* eingetragen. So dauerte es zwar etwas mehr als eine Woche bis alle Abstracts in der *SWEET-DB* kopiert waren, aber die Performance der *PUBMED*-Computer war so zu keiner Zeit gefährdet. Diese Abstracts bildeten dann die Grundlage für die weiter unten beschriebene Klassifizierung der Texte (siehe Abschnitt 5.9).

5.7.11 Ergebnis und Diskussion

Mit den hier vorgestellten Werkzeugen wurde der notwendige zeitliche Aufwand, eine Literatur-Datenbank auf dem Laufenden zu halten, entscheidend reduziert. Dieser Prozess kann zwar noch nicht vollautomatisch durchgeführt werden, da sonst eine zu große Anzahl von nicht relevanten Datensätzen in die Datenbank eingefügt würde. Mit der manuellen Verwaltung einer Liste von Trivialnamen konnte eine sehr gute Ausbeute bei den gefundenen Publikationen erzielt werden. Zu Beginn der automatischen Suche befanden sich 14403 Publikationen in der Datenbank. Mit Hilfe

des Programms *AUTOREFENCE* und einer anschließenden Sichtung eines Experten unter Verwendung des Programms *REFERENCE* konnten 961 neue Datensätze zu den Publikationen hinzugefügt werden. Das heißt, es wurde in einem Zeitraum von weniger als zwei Wochen der Datenbestand an Publikationen um 6,6 Prozent erhöht, und zumindest für die mit einem Trivialnamen identifizierten Strukturen auf den aktuellen Stand gebracht. Dieses ist eine sehr gute Ausbeute, wenn man bedenkt, dass in den beiden ursprünglichen Datenbanken *SUGABASE* und *CARBBANK* die Publikationsdaten manuell gesichtet wurden und auch manuell eingetragen wurden. Im Moment kann die Extraktion der Publikationen zwar nur für Substanzen mit einem Trivialnamen erfolgen, aber ein grundsätzlicher Zugang zur *PUBMED* ist etabliert. Die Vermeidung von Tippfehlern war durch den konsequenten Verzicht von Texteingaben und direkter Übernahme der *PUBMED*-Daten gewährleistet. Bei der weiteren Verarbeitung konnten auch Tippfehler vermieden werden, da der Selektionsprozess durch den Experten durch einfaches Klicken mit der Maus erfolgen kann. Mit Hilfe des Programms *GETABSTRACTS* konnte innerhalb von einer Woche ein Abgleich mit den Datensätzen der *PUBMED* erfolgen und der Datenbestand um die Abstract-Texte ergänzt werden. Die vorgestellten Methoden sind sowohl benutzerfreundlich, als auch effizient.

5.8 Automatische Extraktion von Strukturdaten aus Internet-Quellen

Um den Personalaufwand zur Analyse von Texten möglichst gering zu halten, sollten im Rahmen dieser Dissertation Überlegungen angestellt werden, inwieweit die automatische Extraktion von Strukturdaten aus Internetquellen möglich ist. Dabei kommen zwei Quellen in Frage: Zum einen ist dieses die *PUBMED*, die eine sehr einfache Suche gestattet und auch entsprechende Schnittstellen für die Extraktion des Abstract bereithält, zum anderen sind dieses die Online-Versionen von Journals, die sich thematisch mit Kohlenhydraten beschäftigen wie z.B. ‚Carbohydrate Research' oder aber ‚Glycobiology'.

5.8.1 Anforderungen an die Quelle

Damit es zu einer problemlosen Analyse der Texte kommen kann, müssen einige Anforderungen durch die Datenquelle erfüllt sein:

1. Sie muss frei durch das Internet zugänglich sein.
2. Sie muss automatisch indizierbar sein, damit die Daten sukzessive extrahiert werden können.
3. Die Strukturdaten müssen in Textform enthalten sein, da sonst keine Analyse möglich ist.

Als erstes wurden die oben genannten Quellen überprüft, ob sie diesen Anforderungen genügen:

5.8.2 *PUBMED*

Die *PUBMED*-Datenbank stellt eine definierte Schnittstelle zur Verfügung, die es gestattet einen Eintrag nach dem anderen maschinell herunter zu laden und durch einen Algorithmus analysieren zu lassen. Dieses geschieht relativ einfach durch das Zusammensetzen einer URL, die dann an den *PUBMED*-Server gesendet wird und als Antwort die gewünschten Daten in einer XML-Kodierung zurückliefert. Aus dieser Antwort kann dann ohne Probleme mit der oben beschriebenen Methode zur Analyse von XML-Dateien das Abstract herausgelesen werden. Dies geschieht wie folgt:

> http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=*PUBMED*&retmode=xml&rettype=abstract&id=<n>

wobei <n> der Nummer der *PUBMED*-ID entspricht.

Die *PUBMED*-ID braucht dabei nur hoch gezählt werden, und die URL kann sehr einfach gebildet werden. Die so erhaltenen Abstacts haben weiterhin den Vorteil, dass sie gemäß der XML-Richtlinien[126] in einer UTF-8 Kodierung vorliegen. Alle länderspezifischen Sonderzeichen liegen vor, so dass diese Informationen mit berücksichtigt werden können und nicht verloren gehen.

5.8.3 Carbohydrate Research

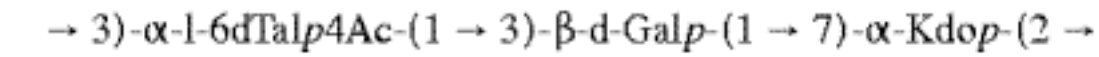
Als nächstes wurde untersucht, inwieweit sich die Online-Version dieses Journals zur automatischen Analyse eignet. Auf die Volltexte kann vom Krebsforschungszentrum aus zugegriffen werden, da die Zentralbibliothek das Journal abonniert hat. Ein Zugriff über das Internet ist also zumindest von einem Rechner des DKFZ aus

möglich. Bei der sukzessiven Adressierung der einzelnen Artikel zeigte sich aber sehr schnell, dass dies von dem Herausgeber der Zeitschrift nicht gewünscht ist. Es bietet sich keine Möglichkeit die URL, wie oben beschrieben, automatisch erzeugen zu lassen, da bei jeder Anfrage ein Code zur Verschlüsselung der übergebenen Daten mit übertragen wird, der sehr wahrscheinlich zur Verschlüsselung des User und einer automatisch erzeugten Session-ID dient. Ist dabei auch nur ein einzelner Buchstabe falsch, führt dieses zu einer Meldung, dass man nicht berechtigt ist, auf diese Informationen zuzugreifen.
Eine Analyse, der manuell erhaltenen Volltexte ergab zudem, dass ein Teil der Struktur-Informationen nicht als Text sondern als Bildinformationen vorliegen. So werden die griechischen Symbole ‚□' und ‚□' nicht alphanumerisch sondern als Bilder dargestellt. Ebenso verhält es sich mit dem ‚D' und dem ‚L' sowie dem Pfeil, der eine kovalente Bindung zwischen zwei Residuen darstellen soll. Ein weiteres Hindernis stellt die unterschiedliche Grammatik dar, die zur Darstellung der Saccharide benutzt wird. So wird in ein und derselben Ausgabe des Journals zum einen die folgende Darstellung gewählt:

□-D-Glc*p*-(1→2)[□-D-Glc*p*-(1→3)]-□-D-Glc*p*-Ome

Darstellung in Iupac-Nomenklatur[127]

oder:



Darstellung als Vollgrafik[128]

sowie:

□-L-rhamnopyranosyl-(1→2)-[□-L-rhamnopyranosyl-(1→4)]-□-D-glucopyranoside

Darstellung als ausgeschiebene IUPAC-Nomenklatur[129]

Anders als bei der Schnittstelle für die *PUBMED* handelt es sich hier nicht um eine für die Öffentlichkeit freigegebene Schnittstelle, so dass schon kleinste Änderungen dazu führen, dass das Analysetool unbrauchbar wird.

5.8.4 Glycobiology

Auch dieses Journal ist in einer Online-Version vom DKFZ abboniert und kann von einem Rechner des DKFZ-Netzwerkes erreicht werden.
Es ergeben sich aber ähnliche Probleme wie bei Analyse der HTML-Seiten von ‚Carbohydrate Research'. So sind auch hier Maßnahmen ergriffen worden, dass es nicht zu einer vollautomatischen Extraktion der Daten kommen kann. Es werden aber unterschiedliche Wege dabei beschritten. Anders als bei der Website von ‚Carbohydrate Research' wird hier die Session-ID nicht bei Aufruf der URL mit gesendet, sondern sie wird als Cookie auf dem jeweiligen Rechner gespeichert. Dieses ist bei Verwendung eines Browsers nicht weiter tragisch. Die Cookies können aber bei Auswertung mittels eines Auswertetools nicht berücksichtigt werden.
Ebenso wie bei dem obigen Journal ‚Carbohydrate Research' werden hier teilweise

für Darstellung der Konfiguration und der Diastereomere Grafiken benutzt, die nicht auszuwerten sind.
Auch in diesem Journal ist keine einheitliche Darstellung der Saccharide gewählt worden. So werden auch hier in einer Ausgabe die folgenden Darstellungen gewählt:

GalNAcß1–4Galß1–3GalNAcß1–4Galß1–4Glc
Darstellung in verkürzter IUPAC-Nomenklatur[130]

Fuc(□1–3)Gal(ß1–4)[Fuc(□1–3)]GlcNAc(ß1–3)GalNAc(ß1–4)
Darstellung in verkürzter Nomenklatur mit anderer Klammersetzung[131]

a-Kdo-(2→8)-a-Kdo-(2→4)-a-Kdo
Darstellung in IUPAC-Nomenklatur[132]

Im Grunde muss für jeden Artikel ein eigener Parser entwickelt werden, der die in dem Artikel enthalten Kohlenhydrate erkennt. Auch hier handelt es sich nicht um eine definierte Schnittstelle, die zu den entsprechenden Problemen führt.

5.8.5 Manuelle Extraktion

Um einen echten Vergleich zwischen den automatischen Methoden und der manuellen Extraktion von Strukturen durch Sichtung der entsprechenden Publikation zu erhalten, musste auch eine manuelle Extraktion durchgeführt werden. Dazu wurden einfach 10 Publikationen eines jeden Journals ausgedruckt und anschließend ausgewertet, und die Daten wurden in die *SWEET-DB* eingetragen. Die Sichtung und die Eintragung in die *SWEET-DB* dauerte für das Journal ‚Glycobiology' 38 Minuten und für die 10 Artikel von ‚Carbohydrate Research' 27 Minuten. Dabei stellte sich heraus, dass der geschwindigkeitsbestimmende Schritt nicht die Sichtung der einzelnen Publikationen war, sondern die entsprechenden Schnittstellen zum Eintragen in die *SWEET-DB* und zum Auslesen *PUBMED* waren. Die weiter oben beschriebenen Tools *GETABSTRACTS,* zur Suche in der *PUBMED,* und *REFERENCE*, zum Eintragen in die *SWEET-DB,* waren dabei eine sehr große Hilfe. In den letzten Jahren sind im Journal ‚Glycobiology' durchschnittlich 96 Artikel und im Journal ‚Carbohydrate Research' durchschnittlich 256 Artikel veröffentlicht worden. Dieses bedeutet, wenn man von den obigen Ergebnissen ausgeht, dass beide Journale innerhalb von etwa 36 Stunden gesichtet und in die *SWEET-DB* eingetragen werden können. Dieses steht in keinem Verhältnis zu der Zeit, die nötig ist, um die Hindernisse beim Zugriff auf die Volltexte auszuräumen und auch bei der Entwicklung der einzelnen Grammatiken, was insgesamt mehrere Wochen gedauert hat.

5.8.6 Ergebnis und Diskussion

Zusammenfassend lässt sich sagen, dass die *PUBMED*, ähnlich wie bei Blaschke et al.[133], als einzige mögliche Quelle für eine Textanalyse in Frage kommt, da nur hier eine automatische Extraktion der Daten möglich ist, und die Daten in einer für Computer verständlichen Form zur Verfügung stehen. Auf Grund der großen Menge von Abstracts lohnt sich hier auch die Entwicklung unterschiedlicher Parsern für die unterschiedlichen Grammatiken, in denen die einzelnen Verbindungen dargestellt werden.

Eine Analyse von kommerziellen Journals beinhaltet einen nicht zu vertretenden Aufwand, da eine manuelle Analyse der Texte zu bedeutend besseren Ergebnissen als eine aufwendige Entwicklung von Textinterpretern und -parsern führt, die dann nur in der Lage wären eine Handvoll Strukturen aus dem jeweiligen Artikel zu extrahieren. Der aufwendige Zugang zu den Webseiten verhindert eine automatische Extraktion. Dieses ist aber nicht so schlimm, da diese beiden Journals von der *PUBMED* indiziert und bei der automatischen Analyse mit erfasst werden und den Nutzern der *SWEET-DB* ebenfalls zur Verfügung gestellt werden.
Auf Grund der großen Menge an Abstracts hält sich die Ausbeute der gefundenen Strukturen in Grenzen. So sind in einem Zeitraum von drei Wochen etwa 23 Strukturen gefunden worden, wobei 13438 Abstracts geparst worden sind. Dieses entspricht einer Ausbeute von 1,71 Promille. Da diese Strukturen allerdings vollautomatisch gefunden worden sind und neu in die *SWEET-DB* eingetragen werden konnten, ist dieser Aufwand allerdings vertretbar. Im Moment muss man leider sagen, dass ich keine echte Alternative zur manuellen Extraktion der Strukturdaten gefunden habe.

5.8.7 Technische Umsetzung

Bei der Umsetzung konnte auf die Klassen zurückgegriffen werden, die schon für die Anwendung *GETABSTRACTS* entwickelt worden sind. Mit Hilfe dieser Klasse ist es möglich, durch einfaches Hochzählen des von der *PUBMED* vergebenen Identifier ein Abstract nach dem Anderen zu übertragen und anschließend mit Hilfe der Grammatiken eine Strukturerkennung zu ermöglichen.

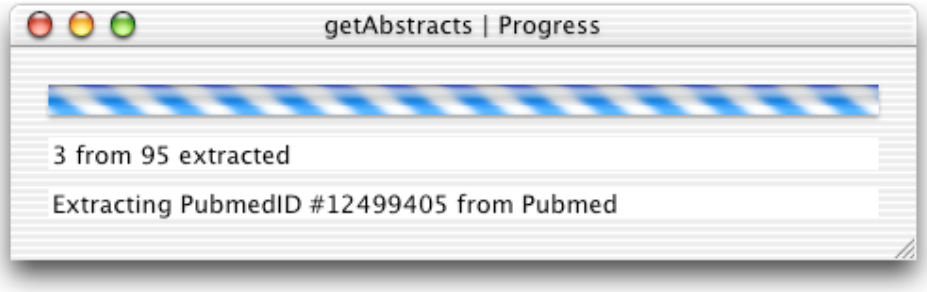


Abbildung 92: Interface des Programmes *GETABSTRACTS*

Die Grammatiken haben im Grunde genommen nur drei verschiedene Teile einer Struktur zu erkennen. Das wichtigste ist das Erkennen der Residuen. Ein Residuum hat vereinfacht geschrieben immer denselben Aufbau

□-D-Glucose

Die Routine zum Erkennen der Residuen musste also in der Lage sein, zu erkennen, ob ein Zucker mit einer □- oder □-Verknüpfung in der D- oder L-Konfiguration vorliegt. Zusätzlich können hier noch weitere Merkmale entsprechend der IUPAC-Nomenklatur für Zucker ermittelt werden.
Ein weiteres Element der Grammatik bestand in dem Erkennen der Linkage-Informationen. Auch dieses gestaltete sich sehr einfach, da sie in der Regel in der folgenden Form vorlagen:

-(1--4)-

Hier wurde von der Grammatik festgestellt, welche Kohlenstoff-Atome an der Bindung beteiligt sind. Mit Hilfe dieser beiden Grundlegenden Funktionen erkennt man nun die beiden wichtigsten Textbausteine einer textlich dargestellten Zuckerstruktur. Durch Klammern werden mehrere Residuen zusammengefasst. Es muss daher auch die korrekte Klammerung erkannt werden. Zum Erkennen einer Zuckerstruktur geht man nun wie folgt vor: Man scannt den Text Zeichen für Zeichen. Sobald man ein Zuckerresiduum erkannt hat, versucht man davon ausgehend, die Parameter der glykosidischen Bindung zu ermitteln und versucht noch weitere Residuen oder Bindungen zu erkennen. Man ist so also ohne Probleme in der Lage, Strukturen zu erkennen, wenn Sie in einem Text vorkommen.
Diese an sich sehr einfachen Regeln ließen sich sehr schnell umsetzen. Probleme bereiteten allerdings wie oben beschrieben die von Publikation zu Publikation unterschiedlichen Darstellungen der Kohlenhydrate. Zumal mindestens ein Tag benötigt wurde, um eine Grammatik zu erstellen und auch zu testen.
Zum Eintragen der Daten in die *AUTO-SWEET-DB* kann nun wieder auf die schon entwickelten Schnittstellen zur *SWEET-DB* zurückgegriffen werden. Als erstes wird die Linucs-Id ermittelt, die als Referenz für alle weiteren Daten in der *SWEET-DB* dient. Dieses geschieht mit der von Alexander Loss entwickelten Schnittstelle, die auch gleichzeitig den korrekten Linucscode erstellt. Die publikationsspezifischen Daten werden von der *PUBMED* ermittelt und können in die Datenbank eingetragen werden. Dieser Vorgang kann vollautomatisch geschehen. Er wird aber im Moment noch von einem Experten kontrolliert, damit keine völlig falschen Einträge in die eigentliche *SWEET-DB* gelangen.
Für eine effiziente Auswertung von Datenbanken, die einen Open-Access Zugang besitzen, sollte daher eine XML-Schnittstelle für die Extraktion der Daten vorhanden sein. Die gespeicherten Datensätze müssen sukzessive adressierbar sein und die Nomenklatur, z.B. für Strukturen, sollte einheitlichen Regeln folgen.

5.9 Semantische Analyse und Klassifizierung von Texten

Ein weiterer Weg Daten aus den Publikationen zu extrahieren besteht darin, manuell Daten aus den Publikationen zu ermitteln. Dieses können zum einen Strukturdaten sein, zum anderen kann es sich aber auch um taxonomische oder biologische Informationen handeln. Ein einzelner Benutzer kann aber nicht einem Abstract nach dem andern in der *PUBMED* durchgehen. Das heißt, es muss auch hier eine manuelle Vorauswahl erfolgen, ob ein Artikel in den Bereich der Glykobiologie gehört oder nicht. Es sollte also im Folgenden untersucht werden, inwieweit es möglich ist, einen Text automatisch in ein Themengebiet einzuordnen.

5.9.1 Statistische Auswertung der Texte

Als erstes wurden alle in der *SWEET-DB* vorhandenen Abstracts genommen und statistisch ausgewertet, welche Wörter darin vorkommen. Dieses führt zu einer sehr interessanten Aufstellung der Wortverteilung der bestehenden *SWEET-DB*. So enthalten die Abstracts, die in der Datenbank enthalten sind, 79450 verschiedene Wörter und Begriffe. Diese Auswertung bildete dann die Basis für die Einordnung der aus der *PUBMED* extrahierten Texte.

5.9.2 Textklassifizierung

Die Semantik eines Satzes kann von nur sehr wenigen Wörtern bestimmt werden, so sind die beiden Sätze:

„Das kleine Kind nähert sich der nahe gelegenen Bushaltestelle“

Und der Satz

„Der kleine Ligand nähert sich der nahe gelegenen Dockingstelle“

zu achtzig Prozent identisch. Dieses legt den Schluss nahe, dass der Sinn eines Textes nur von einer geringen Anzahl von Wörtern bestimmt werden kann, da der Rest zu einem gemeinsamen Grundwortschatz gehört.
Ein neu einzuordnender Text, wenn er in den Bereich der Glykobiologie gehören soll, sollte einen sehr hohen Anteil an den schon gespeicherten Begriffen und Worten enthalten, die sich in den in der *SWEET-DB* gespeicherten Abstracts befinden. Füllwörter, die in jedem englischsprachigen Text vorkommen und damit dem Grundwortschatz entsprechen, wurden bewusst nicht aus der Liste gestrichen, da davon ausgegangen werden kann, dass es diese Worte in jedem Text gibt. Aber ab einem bestimmten Wert, wie es der Beispielsatz zeigt, nur noch für das jeweilige Themengebiet entscheidende Worte zu einem Treffer führen. Dieses bedeutet, dass zwar auch immer vorkommende Begriffe wie ‚that’, ‚the’, ‚a’, ‚these’, und auch ‚an’ zu Treffern führen, aber glykobiologisch relevante Begriffe wie ‚oligosaccharide’ oder ‚trehalose’ die entscheidenden Treffer für die richtige Zuordnung liefern. Für die Klassifizierung wurde nun der Prozentsatz der Wörter ermittelt, die sich schon in den Abstract der *SWEET-DB* befinden.

5.9.3 Ermittlung eines Grenzwertes für die Klassifizierung

Als erstes wurde ein Grenzwert ermittelt, wann ein Text in den Themenbereich der *SWEET-DB* einsortiert werden kann. Dazu wurden zwanzig Abstracts aus der

PUBMED manuell klassifiziert. Es wurden 10 Texte ausgesucht, die eindeutig in den Bereich der Glykobiologie gehören und 10 Texte, die eindeutig nicht dazu gehören. Zuerst wurden die 10 Texte, die eindeutig in das Themengebiet gehören, analysiert. Die entsprechenden Ergebnisse sind in der Tabelle dargestellt:

Pubmed ID	Häufigkeit (in Prozent)
10536041	93,310
10536039	98,537
10536038	98,701
10536037	95,785
1964872	88,889
1964871	86,429
1964870	95,082
1964868	95,455
1964867	97,753
10536036	94,545

Tabelle 17: Klassifizierung der in das Themengebiet gehörenden Texte

Anschließend wurden die 10 fachfremden Texte analysiert. Die Ergebnisse sind in der folgenden Tabelle dargestellt:

Pubmed ID	Häufigkeit (in Prozent)
9849192	91,176
2124893	80,263
2124887	82,222
11081558	90,385
1854628	91,964
1673430	81,119
1563422	81,818
1778887	85,714
7672866	91,324
8787376	90,526

Tabelle 18: Klassifizierung der nicht in das Themengebiet gehörenden Texte

Der Grenzwert für die Klassifizierung eines Textes zu einem Themengebiet wurde empirisch ermittelt. Die Grenze wurde so festgelegt, dass die gefundenen Artikel sicher in den Bereich der Glykobiologie gehören. In der folgenden Tabelle sind die Ergebnisse der Zuordnungen in Abhängigkeit von dem Grenzwert zu sehen.

Grenzwert	Richtig eingeordnete Texte	Falsch positive	Falsch negativen
88,0	14	5	1
89,0	13	5	2
90,0	13	5	2
91,0	15	2	0
92,0	18	2	0
93,0	18	2	0
94,0	17	3	0

Tabelle 19: Ergebnisse der empirischen Ermittlung des Grenzwertes

Anhand dieser Tabelle wurde der Grenzwert für eine richtige Zuordnung auf 91 festgelegt, da mit diesem Grenzwert die beste Klassifizierung der Texte erfolgt ist. Vor allen Dingen wurden keine Texte falsch in den Bereich der Glykobiologie eingeordnet.

5.9.4 Technische Umsetzung

Die Umsetzung dieser Methode gestaltete sich sehr einfach, da die meisten Funktionen schon in vorhergehenden Projekten entwickelt worden sind. Zum einen war es nötig, dass die in der *SWEET-DB* enthaltenen Abstracts statistisch ausgewertet werden. Dazu wurde ein Programm entwickelt, dessen einzige Aufgabe darin bestand, alle gespeicherten Abstract nacheinander durchzugehen und die gefundenen Worte in eine Datenbank zu schreiben. Die zu speichernden Daten bestanden nur aus zwei Feldern:

1. dem Wort
2. der Anzahl der Worte

Dabei wurde das Wort selbst als Index verwendet, und es kann so durch eine einfache SQL-Abfrage festgestellt werden, ob und wie oft ein Wort in den Texten der gespeicherten Abstracts vorkommt. Dabei wurde nicht zwischen Groß- und Kleinschreibung unterschieden. Außerdem wurden sämtliche Satzzeichen ignoriert, um Wörter nicht unnötigerweise zwei Einträgen zuzuordnen.
Wie nicht anders zu erwarten, waren die am häufigsten benutzten Worte, die Worte ‚the', ‚a', ‚these' und ‚at'. Das am häufigsten benutzte Wort in der Glykobiologie war das Wort ‚oligosaccharide', das 3336 mal vorkam. Die statistische Häufigkeit eines Wortes wird zurzeit aber nicht berücksichtigt, da im Moment nur eine qualitative Suche Berücksichtigung findet.
In einem zweiten Programm wird ein Score ermittelt, der festgelegt, wie hoch der prozentuale Anteil der Wörter ist, die in den Texten der *SWEET-DB* vorkommen. Der prozentuale Anteil der Wörter lässt sich sehr leicht mit Hilfe der folgenden Formel ermitteln:

$$\text{Häufigkeit}_{\text{Prozent}} = \frac{\text{Anzahl}_{\text{dergefundenenWörter}}}{\text{Gesamtanzahl}_{\text{derWörter}}} * 100$$

Ist dieser Score größer als der oben ermittelte Grenzwert, so gehört der Text in den Bereich der Themen, die in den in der *SWEET-DB* beschriebenen Publikationen behandelt werden.

5.9.5 Ergebnis

Um die Qualität dieser Methode festzustellen, wurden nun wie oben beschrieben 40 zufällige Texte genommen und von dem obigen Programm ausgewertet und anschließend zugeordnet. Dieses führte zu folgendem Ergebnis:

Kategorie	Anzahl
Richtig zugeordnete Texte	38
Falsch positiv zugeordnete Texte	0
Falsch negativ zugeordnete Texte	2

Tabelle 20: Klassifizierung der Texte

Die Methode führt zu einem im Verhältnis zum Aufwand sehr befriedigendem Ergebnis. Es gibt eine sehr geringe Fehlerquote. Kein Text, der nicht in das Themengebiet gehörte, wurde eingeordnet.

5.9.6 Entwicklung eines Testsystems auf Vollständigkeit

Ein wichtiges Kriterium für ein Klassifizierungssystem ist die Vollständigkeit. So sollten auch wirklich alle Texte, die in den Bereich der Glykobiologie gehören, von diesem System auch erkannt werden. Dieses lässt sich am einfachsten überprüfen, ob zum Beispiel alle Artikel eines Jahrgangs der Zeitschrift ‚Glycobiology' von diesem Klassifizierungssystem erkannt werden. Eine Recherche für das Jahr 2002 in der *PUBMED*-Datenbank ergab 96 Artikel, die in diesem Zeitraum in dieser Zeitschrift veröffentlicht worden sind.
Nachdem die Abstracts mit Hilfe des Programms *GETABSTRACTS* aus der *PUBMED*-Datenbank ausgelesen worden sind, wurden diese Abstracts mit Hilfe der Webschnittstelle klassifiziert. Von den 96 Artikeln wurden 88 in den Bereich der Glykobiologie eingeordnet und sie so hätten ausgewertet werden können.

5.9.7 Diskussion und Ausblick

Mit den dargestellten Mitteln konnte dargelegt werden, dass bei Vorhandensein einer schon mit in ein Themengebiet eingeordneten Menge an Texten, es relativ einfach ist, Texte in ein Wissensgebiet einzuordnen, um diese dann anschließend weiterzuverarbeiten. Mit den vorgestellten Methoden wird eine ausreichende Vollständigkeit, in der Regel über 80%, und Richtigkeit der Zuordnung erreicht. Man braucht nur eine gewisse Menge an Texten aus einem Themengebiet nehmen und kann dann neue Texte durch ein Parsen, ähnlich wie bei der automatischen Extraktion von Strukturinformationen, sehr einfach diesem Wissensgebiet hinzufügen. Ein weiterer Vorteil dieser Methode besteht darin, dass man die Daten in einem Format erhält, das sehr einfach weiterverarbeitet werden kann. Leider konnte nicht untersucht werden, inwieweit eine Vorauswahl der Texte das Ergebnis beeinflusst hätte. Es befinden sich zurzeit etwas weniger als 10.000 Abstracts in der *SWEET-DB*. Viele Texte in der *SWEET-DB* sind aus dem Bereich der pharmazeutischen Biologie, so dass sich in dem Vokabular auch viele Begriffe aus diesem Themengebiet[134-136] befinden. Diese Texte sind sehr schwierig von anderen Texten der Glykobiologie abzugrenzen, da sich sehr ähnliche Begriffe darin befinden. Sicherlich ließe sich das Klassifizierungssystem noch verbessern, wenn man diese Texte aus der Datenbank löschen würde.

6 Zusammenfassung

Die vorliegende Arbeit umfasst die Entwicklung von Algorithmen und Strategien zur Analyse von Massenspektren von Glykanen sowie Strategien zur Aktualisierung und Annotierung einer bestehenden Datenbank, der *SWEET-DB*.
Für die Glykobiologie fehlte es bisher an Algorithmen, die ähnlich wie im Bereich der Proteomik bei der Sequenzierung von Peptiden, dem Benutzer eine Hilfe bei der Analyse von N-, O-Glykanen und Lipopolysacchariden sind. Die Zusammensetzung dieser Verbindungen ist aber für das Verständnis der zellulären Stoffwechsel-physiologie von essentieller Bedeutung.

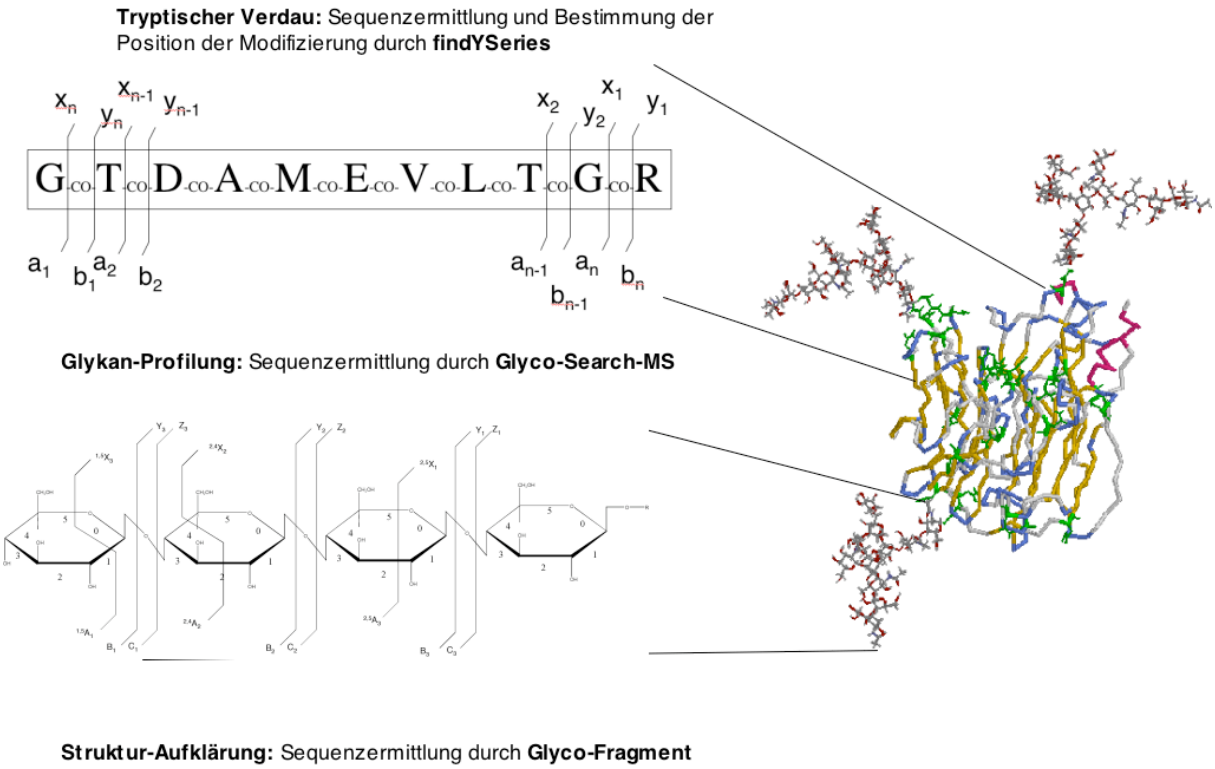


Abbildung 93: Einsatzgebiete der Algorithmen

Im Rahmen der Entwicklung von Algorithmen zur Aufklärung von Massenspektren wurden insgesamt drei Programme entwickelt, die es dem Forscher gestatten, eine große Anzahl von Spektren, die im Bereich der Proteomik und Glykomik anfallen, auszuwerten.

1. ***FINDYSERIES***: Ausgehend von einem MS/MS-Spektrum eines mit Trypsin verdauten Proteins oder Peptids wird untersucht, ob und wo sich eine co- oder posttranslationale Modifikation der Peptidsequenz befindet. Der entwickelte Algorithmus basiert auf einen direkten Zugriff auf das Softwarepaket *MASCOT*, mittels dem die Peptidsequenz ermittelt wird.
2. ***GLYCO-FRAGMENT***: Mittels dieser Web-Applikation ist es möglich, interaktiv alle Fragmente von vorgegebenen Kohlenhydraten generieren zu lassen auf der Grundlage der Fragmentierungsregeln von Domon und Costello. Die so

erhaltenen Fragment-Ionen können dann mit den experimentell gemessenen Massen verglichen werden.

3. ***PEAKASSIGN***: Mit dieser Web-Applikation ist eine komfortable Möglichkeit geschaffen worden, die in einem Massenspektrum enthaltenen Ionen entsprechend der Nomenklatur von Domon und Costello zuzuordnen. Zusätzlich zu *GLYCO-FRAGMENT* wird hier ein Teil der inneren Fragment-Ionen berechnet und zugeordnet. Als Eingabe werden das zu analysierende MS-Spektrum und ein Strukturvorschlag benötigt.

Mit diesen Programmen wurde Möglichkeiten zur schnellen Interpretation von MS-Spektren geschaffen, mittels derer eine große Anzahl von typischen Fragestellungen im Bereich der Analytik von Glykanen bearbeitet werden können. So kann mit diesen Programmen die Komposition, die Anzahl der Antennen oder in günstigen Fällen sogar die komplette Sequenz eines Kohlenhydrats ermittelt werden.
Im dem Maße wie durch effiziente Algorithmen die Auswertung und Zuordnung von Messdaten vereinfacht und optimiert wird, steigt auch die Menge der daraus gewonnenen interpretierbaren Daten, Informationen und Erkenntnisse. Da die so entstehenden Daten zumeist schon in digitaler Form vorliegen, ist es sinnvoll, der wissenschaftlichen Allgemeinheit diese in entsprechenden Datenbanken zur Verfügung zu stellen. In der Vergangenheit beruhte der Prozess der Bereitstellung von Daten und Informationen zumeist auf manueller, nachträglicher Extraktion von Daten aus Publikationen. Sie war deshalb durch einen hohen Aufwand an menschlicher Arbeit gekennzeichnet. Gerade im Bereich der Glykowissenschaften hat es sich leider gezeigt, dass dieser Prozess durch die damit verbundenen Kosten beim Auslaufen der staatlichen Förderungen schnell zum Ende eines Projektes führen kann. In dieser Arbeit sind verschiedene Strategien zur Pflege einer Datenbank entwickelt worden. Sie wurden im Hinblick auf ihre Potenz untersucht, eine automatische Annotierung der Daten zu gestatten. Bei der Umsetzung sind zwei Erweiterungen zur Identifizierung von Kohlenhydrat-Strukturen der *SWEET-DB* entstanden.

1. ***GLYCO-SEARCH-MS***: Hier wurde basierend den in der *SWEET-DB* gespeicherten 12.000 Strukturen von Glykanen mit Hilfe des *GLYCO-FRAGMENT*-Algorithmus eine Spektrenbibliothek von theoretisch berechneten Fragmenten (A, B, C, X, Y, Z) geschaffen. Diese berechneten Spektren werden mit Peaklisten gemessener Massenspektren verglichen. Die implementierte Bewertungsfunktion gibt dem Benutzer ein geeignetes Kriterium an die Hand, um entscheiden zu können, ob ähnliche oder identische Strukturen in der *SWEET-DB* vorhanden sind. Dabei werden die im Bereich der Glykomik eingesetzten massenspektroskopischen Methoden unterstützt. Das verwendete Esi-Ion kann ebenso frei gewählt werden wie die verwendete Signalart und auch der Toleranzbereich für die Erkennung von Treffern bei der Suche in der Spektrenbibliothek.
2. ***GLYKAN-PROFILING***: Aufbauend auf die obige Spektrenbibliothek wurde eine schnelle Identifizierung von Strukturen an Hand des Molekülpeaks geschaffen. Auch hier lassen sich die Messparameter entsprechend einstellen.

Bei der Neueingabe von Daten wurden zwei unterschiedliche Strategien entwickelt, die so effizient gestaltet wurden, dass sie für die Routine-Eingabe durch Wissenschaftler aber auch Studenten, die über eine entsprechende Vorbildung verfügen, gut geeignet sind.
Bei der Neueingabe von Daten sind in erster Linie die **Arbeitsumgebung zur Verwaltung von NMR- und Massenspektren** zu nennen. Es wurde eine dezentrale Lösung geschaffen, die es dem Benutzer ermöglicht, seine lokal gemessenen Spektren in dieser Datenbank zu verwalten. Hat er seine Ergebnisse veröffentlicht, können die Spektren über die beschriebenen Schnittstellen sofort in der *SWEET-DB* allgemein zugänglich gemacht werden. Dieses Vorgehen hat den Vorteil, dass die Daten ohne erneute Eingabe in die Datenbank übernommen werden können. In einem ersten Test wurden von zwei Studenten innerhalb von 6 Monaten 347 Spektren aus der Literatur exzerpiert und über die Arbeitsumgebung eingegeben, und sie stehen nun der *SWEET-DB* zur Verfügung.
Mit Hilfe der Programme ***AUTOREFERENCE*** und ***REFERENCE*** konnte eine Aktualisierung der Literatureinträge von in der *SWEET-DB* bereits vorhandenen Strukturen, die durch einen Trivialnamen charakterisiert sind, semiautomatisch erfolgen. Ausgehend von einer Liste mit Trivialnamen kann in regelmäßigen Abständen in der *PUBMED* gesucht werden. Diese Rohdaten werden in einer temporären Datenbank zwischengespeichert und werden nach einer Kontrolle durch einen Experten in die *SWEET-DB* überführt.
In Zusammenarbeit mit Mitarbeitern des DKFZ entstanden zwei weitere Anwendungen, die thematisch nicht der Glykobiologie zuzurechnen sind. Sie ließen sich jedoch mit den im Rahmen dieser Arbeit entwickelten Algorithmen und Software Werkzeugen gut bearbeiten.

1. Mit dem Programm ***AUTOMASCOT*** wurde eine Möglichkeit geschaffen, routinemäßig anfallende Massenspektren von digestierten Proteinen unter Verwendung des *MASCOT*-Softwarepaketes automatisch die Sequenz zuzuordnen. Die Ergebnisse können über ein Webinterface dargestellt werden, und es können so auch Spektren ausgewertet werden, für die bei einer manuellen Auswertung die Zeit fehlt. (Zusammenarbeit mit Prof. Lehmann, DKFZ, Zentrale Spektroskopie)

2. Es wurde ein **Webinterface zur Vorhersage der Selektivität eines Substrats für die COX-II** entwickelt. Damit kann über das Internet unter Verwendung des Programmpakets *AUTODOCK* eine virtuelle Testung von Verbindungen erfolgen, und im Falle einer möglichen Präferenz für die COX-II können weitere in-vitro Untersuchungen unternommen werden. (Zusammenarbeit mit Prof. Werner)

Bei der Realisierung dieser Projekte wurden zwei verschiedene Strategien bei der Implementierung der Algorithmen verfolgt. Zum einen wurde der klassische Ansatz eines zentralen Servers mit Darstellung der Ergebnisse in einem Browser verwendet. Dieser wurde verglichen mit dem im Moment favorisierten Ansatz einer Einzelplatz-Anwendung, die nur den Austausch der Daten über das Internet vornimmt. Die Berechnungen und Darstellung der Ergebnisse aber erfolgen auf dem lokalen Rechner. Dabei konnte gezeigt werden, dass für einfachere Aufgaben und Darstellungen der erste Ansatz ausreicht. Für aufwendigere Aufgaben, die eine

schnelle Interaktion mit dem Benutzer und eine aufwendige grafische Darstellung der ermittelten Ergebnisse erfordern, ist eindeutig die zweite Lösung vorzuziehen.

7 Ausblick

7.1 Entwicklung von Algorithmen für die Massenspektrometrie

Mit den hier vorgestellten Algorithmen und Strategien zur Auswertung von Massenspektren im Bereich der Glykobiologie sind grundlegende Arbeiten geleistet worden für eine effiziente, computergestützte Identifikation von Glykanstrukturen bzw. deren Komposition. Momentan werden weltweit verschiedene Glykomik-Projekte durchgeführt und weitere initiiert, in denen die Massenspektrometrie zur Bestimmung der Glykan-Profile intensiv verwendet wird. Eine entsprechende Softwareunterstützung wie bei den Proteinen, wo sich die *MASCOT*-Software als quasi Standard etabliert hat, gibt es für die Identifizierung der Glykane bisher nicht. Mit den in dieser Arbeit beschriebenen Algorithmen und Strategien wurden jedoch alle notwendigen elementaren Bausteine für die Entwicklung einer entsprechenden Software gelegt. Aus diesem Grund sollen in Zukunft Anstrengungen in diesem Bereich verstärkt werden. Insbesondere ist es nötig, die Programme zur Auswertung der Massenspektren noch weiter zu automatisieren, so dass diese Programme routinemäßig und vollautomatisch im Batch-Betrieb eingesetzt werden können. Um einen hohen Durchsatz für den Routinebetrieb zu gewährleisten, ist es weiterhin notwendig, die verwendeten Algorithmen im Hinblick auf ihre Effizienz zu optimieren.

7.2 Dezentrale Eingabemöglichkeiten von Spektren

7.2.1 Massenspektren

Mit den beiden aktuell schon verfügbaren Web-Applikationen *GLYCO-SEARCH-MS* und *GLYKAN-PROFILING* sind erste Möglichkeiten geschaffen worden, den momentan existierenden Mangel an Softwarewerkzeugen zur Interpretation von gemessenen Spektren zu beheben. Nach wie vor besteht jedoch das Problem, dass es keine gut validierte Sammlung von Referenzspektren von Glykanen gibt. Die hier vorgestellte Methode zur Berechnung von Fragment-Ionen kann diesem ungünstigen Zustand nur teilweise entgegenwirken. Es ist deshalb eine wichtige Aufgabe, in Zukunft diese Spektren mit real gemessenen Spektren zu ergänzen, da nur so die exakte Identifizierung von Glykanen durch Ihre Vergleichsspektren möglich ist. Die in dieser Arbeit entwickelten, allgemein verfügbaren, dezentralen Eingabemöglichkeiten von Massenspektren sowie die Arbeitsumgebungen zur Verwaltung der Spektren sollen Massenspektroskopiker dazu ermuntern, die von ihnen gemessenen Spektren direkt mittels der hier vorgestellten Werkzeuge zu interpretieren, Massen zuzuordnen und die Spektren zu archivieren. Es besteht die Hoffnung, so den Datenbestand an experimentell gemessenen Massenspektren kontinuierlich zu erhöhen. Erste konkrete Erfahrungen zeigen, dass dieses ein Prozess sein wird, der viel Geduld, eine hohe Flexibilität auf die jeweiligen Bedingungen bei den ‚Kunden' zu reagieren und einiges an Überzeugungsarbeit beinhalten wird.

7.2.2 NMR-Spektren

Die Aussagen für die Notwendigkeit der Speicherung der Originaldaten von Massenspektren von Glykanen, gelten in vollem Umfang auch für die Aufnahme von NMR-Spektren. Vereinfacht wird die Situation bei den NMR-Spektren jedoch dadurch, dass zugeordnete chemische Verschiebungen als Listen von Resonanzen in den Publikationen dokumentiert sind. Damit eröffnet sich die Möglichkeit der

nachträglichen manuellen Extraktion dieser Daten. Auch wenn dieses Verfahren mehrere unnötige Medienbrüche beinhaltet und deshalb schon recht fehleranfällig ist. So war es aktuell die einzige Möglichkeit, dem Mangel an zugeordneten NMR-Spektren entgegen zu wirken. In Zukunft müssen jedoch verstärke Anstrengungen unternommen werden, die NMR-Spektroskopiker davon zu überzeugen, dass die direkte Eingabe und Zuordnung ihrer Spektren in eine Datenbank langfristig Vorteile für sie hat.

7.3 Automatische Aktualisierung der *SWEET-DB*

Die hier vorgestellten Untersuchungen haben ergeben, dass die semi-automatische Aktualisierung der *SWEET-DB* sehr effizient gestaltet werden kann, wenn eine eindeutige Zuordnung zwischen einer Glykanstruktur, die bereits in der *SWEET-DB* enthalten ist, und einem ‚Trivialnamen' wie etwa GM1, GD1a, Lewis^X existiert. Diese bereits erfolgreich praktizierte Strategie soll in Zukunft intensiviert werden, in dem systematisch nach weiteren Beschreibungen von Glykanstrukturen gesucht werden soll, die in Publikationen verwendet werden. Ist dieses nicht möglich, sollten die Versuche, Strukturdaten aus den Publikationen zu extrahieren, intensiviert werden.

7.4 Automatische Erkennung von themenrelevanten Publikationen

Die vorgestellten Werkzeuge zur automatischen Klassifizierung von Texten sollten dazu eingesetzt werden, Literaturstellen für eine neu aufzubauende Glykom-Datenbank zu liefern, die entsprechend den vorgegebenen Kriterien ausgewertet und eingeordnet werden können. Diesen Literaturstellen werden dann zusätzliche Informationen zu den gefundenen Glykanen und den glykosylierten Proteinen hinzugefügt. Aufgrund der beschriebenen Erfahrungen bei der automatischen Extraktion von Glykan-Strukturen aus Originalarbeiten wird eine effiziente manuelle Eingabe von Strukturen notwendig sein. Es ließen sich so Beschreibungen der kompletten Glykome von einzelnen Spezies, Zellinien oder Proteinen erstellen, die es dann erlauben, abgelegt in einer sinnvoll organisierten Datenbank, Vergleiche zwischen den Glykanmustern von gesunden und kranken Zellen durchzuführen. So wäre eine sicherere Identifizierung der Unterschiede in den Glykan-Strukturen möglich. Dieses ist besonders interessant für die Analyse der Glykome von Knockout-Mäusen. Bei diesen Knockout-Mäusen handelt es sich um Mäuse, bei denen einzelne Gene ausgeschaltet sind. Sind durch diese Gene kohlenhydrat-aktive Enzyme kodiert, so können aus dem Glykom Rückschlüsse auf die Metaboliserung von Kohlenhydraten gezogen und Angriffspunkte für neue Therapien erkannt werden.

7.5 Open Access

Die dezentralen Arbeitsumgebungen bilden eine ideale Basis für zukünftige Open Access Projekte. Sie bietet Forschern die Möglichkeit ihre Primärdaten zu verwalten und im Falle einer Publikation stehen diese sofort der Allgemeinheit zur Verfügung. Es gehen so auch nicht mehr wie bei der Veröffentlichung über einen Verlag die Primärdaten verloren, sondern stehen weiterhin zur Verfügung. Dies kann sogar soweit gehen, dass die Original-Messdateien, wie z. B. die Kurvenzüge einer NMR-Messung als JCAMP-DS Dateien, dort abgelegt werden können und jederzeit herunter geladen werden können.

8 Verwendete Technologien

8.1 Apache

Ein Freeware Web Server, der bei über der Hälfte aller Websites eingesetzt wird und auf dem NCSA -Web-Server basiert. Mit viel Funktionalität, einfachem modularen Aufbau und einer Plattformvielfalt vom kleinen 386er bis zur großen Multiprozessor-UNIX-Maschine hat er innerhalb kürzester Zeit mit dem höchsten Anteil aller Web-Server viele Konkurrenten auf die hinteren Plätze verwiesen. Das Konkurrenzprodukt zu Apache ist der IIS. Der Apache kann relativ schnell um verschiedene Skriptsprachen wie PHP oder Perl erweitert werden. Es besteht so die einfache Möglichkeit auch dynamische Webseiten zu erzeugen.

8.2 Cocoa

Bei Cocoa handelt es sich um eine Entwicklungsumgebung für Mac OS X, mit der es möglich ist, auf sehr einfache und elegante Art und Weise Anwendungen dafür zu schreiben. Als Programmiersprache findet hauptsächlich Objectice C Verwendung, wobei es sich um eine auf C basierende Sprache handelt, die um objektorientierte Syntax-Elemente erweitert worden ist. Die Entwicklungsumgebung besteht aus zwei Teilen: Zum einen besteht sie aus einem *PROJECT BUILDER*, mit dem der Sourcecode bearbeitet werden kann, zum andern wird ein *INTERFACE BUILDER* mitgeliefert, mit dem es möglich ist das Userinterface zu gestalten. Die einzelnen Elemente des User-Interfaces sind direkt den Klassen des Cocoa-Frameworks zugeordnet, so dass sämtliche relevanten Parameter direkt eingestellt werden können.
Im Rahmen dieser Arbeit sind eine große Anzahl von Programmen und Tools mit dieser Umgebung entwickelt worden. Das größte Projekt war die Entwicklung des Programms *FINDYSERIES*, aber auch dies gestaltete sich sehr einfach und effizient. Durch die Verwendung der objektorientierten Elemente ist es möglich, entwickelte Klassen immer wieder in Programmen zu verwenden, ohne auch nur eine Zeile Code neu einzutippen. So wurde nur einmal eine Klasse entwickelt, um in der *PUBMED* zu suchen und Daten zu extrahieren. Diese Klasse wurde dann in den Programmen *GETABSTRACT* und *REFERENCE* ohne Änderung verwendet. Auf den Webseiten der Firma Apple[1] befinden sich sehr viele Beispiele für die Gestaltung und Entwicklung von Programmen.

8.3 Javascript

Am Anfang des Internets bestand das Internet nur aus HTML-Dokumenten, die von einem Browser, z. B. dem *NETSCAPE NAVIGATOR* oder dem *INTERNET EXPLORER*, dargestellt werden können. Die einzige Möglichkeit eine Interaktivität mit dem Benutzer herstellen zu können, bestand in der Möglichkeit mit einem CGI-Skript dynamische Seiten zu erstellen. Dazu musste aber jedes Mal eine Anfrage an den Server gestellt werden, deren Ergebnis dann von dem Browser dargestellt werden muss. Dieser Prozess benötigt auf der einen Seite Bandbreite, die der Server zur Verfügung stellen muss, und es kann relativ lange dauern, wenn zur selben Zeit viele Anfragen an den Server gestellt werden, da dieses Skript für jede Anfrage aufgerufen werden muss. Mit der Entwicklung von Javascript wurde dieser Mangel

[1] http://developer.apple.com

beseitigt, da es von nun möglich war die Darstellung eines angeforderten HTML-Dokumentes auch auf der Client-Seite zu ändern. Das Prinzip der Programmiersprache besteht darin, dass jedes dargestellte Element der Seite als ein Objekt aufgefasst wird, das mittels zugehöriger Methoden und Eigenschaften analysiert und verändert werden kann. So ist es möglich Formularfelder zu verändern oder hinzuzufügen.
Außerdem kann auf Events, wie das Beenden des Ladens der Seite reagiert werden. Aber auch auf Aktionen des Benutzers, wie das Anklicken eines Links oder eines Bildes, kann durch z.B. das Austauschen eines Bildes reagiert werden. Die Darstellung der Seite erfolgt bedeutend schneller, da es nicht mehr nötig ist, mit dem Server zu kommunizieren, sondern die Auswertung erfolgt lokal auf dem Rechner des Clients.

8.4 Linux

ist eine frei kopierbare Variante des Betriebssystems UNIX, die als Alternative auf IBM-kompatiblen PCs ab 80386- Prozessor läuft. 1991 begann der finnische Student Linus Torvalds mit der Entwicklung. Die erste öffentliche Linux-Version 0.01 erschien am 17.09.1991 (500 KByte Quelltexte, verteilt auf 10.000 Codezeilen). Seit etwa 1992 wird Linux von einer Reihe von Programmierern in Zusammenarbeit mit Torvalds weiterentwickelt. Inzwischen programmieren rund um die Welt Tausende daran, da die Quellcodes gleich mitgeliefert werden. Im März 1994 gab Torvalds die Version 1.0 frei. Das Betriebssystem ist kostenlos und sehr stabil. Es gibt bereits zahlreiche kostenlose Software-Version für Linux. Linux enthält alle Bestandteile für einen Internet -Server. Es organisiert sehr zuverlässig die User-Rechte.
Linux ist neben dem Apache-Server eines der Vorzeigeprodukte der Open Source - Bewegung. Der Siegeszug begann, als Linus Torvalds am 09.06.1996 Kernel 2.0 auf die offiziellen FTP -Server stellte.

8.5 Mac OS X

Im Jahre 2001 wurde von Apple ein auf BSD Unix basierendes Nachfolge-betriebssystem für das in die Jahre gekommene Mac OS 9.x vorgestellt. Mac OS X verbindet als erstes Betriebssystem die Vorteile einer ausgefeilten grafischen Benutzeroberfläche mit den Stärken eines UNIX-Betriebsystems. Es bietet für das Netzwerk Funktionen, wie WebServer, MySQL-Server und für den einzelnen Benutzer eine gute Officefähigkeit. Seine wirklichen Stärken konnte es aber in dieser Arbeit ausspielen: Es sind dies die mitgelieferten Entwicklungstools. So können mit Hilfe der mitgelieferten Frameworks (unter anderem Cocoa) und der Entwicklungsumgebungen ohne Probleme Programme entwickelt werden, die über ein einheitliches Benutzerinterface verfügen. Eine ähnliche Strategie wird im Moment mit der .NET-Technologie von Microsoft versucht. Bestehende C-Software kann ohne Probleme in die Programme integriert werden.

8.6 Microsoft Windows

Als ‚Windows' bezeichnet man eine grafische Benutzeroberfläche der Firma Microsoft für IBM-kompatible Rechner. MS-Windows war ursprünglich eine Betriebssystem-Erweiterung für MS-DOS und diente als Arbeitsumgebung für spezielle dafür entwickelte Windows-Programme.
DOS-Programme können von Windows aus aufgerufen werden. Windows 1.0 wurde 1985 für den 8086- Prozessor entwickelt und wurde überwiegend zum Starten von

DOS-Anwendungen verwendet. 1987-88 wurde mit Windows 286 und Windows 386 eine erste Anpassung an die neuen Prozessoren von Intel durchgeführt, ohne eine wesentliche Funktionserweiterung. Erst mit Windows 3.0 (1990) wurde der große Erfolg von Windows und Windows-Anwendungsprogrammen eingeleitet, der mit Windows 3.1 und Windows für Workgroups 3.11 ausgebaut wurde. Die wichtigste Neuerung bestand unter anderem in der besseren Ausnutzung der neuen Intelprozessoren und der Verwendung des erweiterten Arbeitsspeichers für alle Programme.
Mit der Version Windows 95 wurde das Programm zu einem eigenen Betriebsystem mit fortschrittlicher Bedienerführung weiterentwickelt. Charakteristik: Kennzeichnend für Windows ist eine einheitliche, standardisierte Steuerung über Symbole, Menüs und graphische Dialogfelder, die überwiegend mit einer Maus bedient werden. Dadurch entfällt die von DOS bekannte manuelle Eingabe von Befehlen. Die Bezeichnung Windows (Fenster) ergibt sich aus der Verwendung von Fenstern für die Darstellung der Arbeitsoberfläche von Anwendungsprogrammen und Dokumenten. Allgemeine Aufgaben wie etwa Drucken sowie die Steuerung von Rechnerkomponenten (z. B. Festplatte, Grafikkarte) werden von MS-Windows zentral verwaltet und allen Anwendungsprogrammen zur Verfügung gestellt. Windows erlaubt ein einfaches Multitasking, also den quasi gleichzeitigen Betrieb von mehreren Programmen.
Wesentlich zur Verbreitung hat auch der vereinfachte Datenaustausch zwischen verschiedenen Anwenderprogrammen beigetragen. Texte und Grafiken können einfach über die so genannte Zwischenablage ausgetauscht werden. Mit Windows 3.x wurden außerdem proprietäre Standards zum interaktiven Datenaustausch definiert, nämlich DDE und OLE. MS-Windows 3.x wurde die am weitesten verbreitete graphische Benutzeroberfläche für Intel-PC's.

8.7 MySQL

MySQL ist eine echte Multi-User, Multi-Treaded SQL-Datenbank und wird von vielen großen Providern oder auch Suchmaschinenbetreibern eingesetzt. MySQL ist eine Client/Server Implementierung, die aus einem Server-Dämon ‚mysqld' und vielen Client Programmen, sowie Bibliotheken für die Programmiersprachen PERL, PHP und ASP besteht.
Structured Query Language (SQL) ist eine standardisierte Datenbanksprache, die das Speichern, Updaten und den Zugriff auf Informationen erleichtert. Beispielsweise kann man Produktinformationen eines Kunden auf einem WWW-Server speichern und abrufen. MySQL ist äußerst schnell und flexibel genug, um sogar Bilder und Log-Dateien darin abzulegen. In dieser Arbeit sind alle Datenbanken mit dieser Datenbank angelegt worden. Dabei ist es nie zu irgendwelchen Datenverlusten oder Abstürzen des Server-Dämons gekommen. Für den Bildungsbereich ist die Nutzung der Datenbank kostenlos.

8.8 PHP

PHP (PHP: Hypertext Preprocessor) ist eine serverseitig interpretierte, in HTML eingebettete Skriptsprache. Die Syntax ist ähnlich zu C, Java und Perl, und erweitert durch PHP-eigene Features wie z.B. Kommandos zur Integration von Datenbanken. PHP gibt dem WWW-Anwendungs-Entwickler ein einfach erlernbares und gleichzeitig mächtiges Werkzeug zur Erstellung von Web-Seiten dynamischen Inhalts an die Hand. PHP existiert sowohl für Unix als auch für die Windows 95/98/NT

Plattform. Sehr beliebt ist die Integration von PHP als Modul in den Apache-Webserver, des Weiteren ist die Ausführung via CGI möglich. In dieser Arbeit sind alle über das Internet erreichbaren Applikationen in PHP entwickelt worden. Die Sprache bietet den Vorteil, dass für fast alle Programmierungsprobleme Funktionen zur Verfügung stehen. Dies war von großem Vorteil für die Erstellung der Spektren-Darstellungen. Hier konnte einfach auf die Grafik-Funktionen zurückgegriffen werden.

9 Anhang

9.1 Weitere Arbeiten

Im Rahmen meiner Dissertation am DKFZ entstanden in der Zusammenarbeit mit Projektpartnern weitere Programme und Webapplikationen, die sich thematisch nicht in die vorherigen Kapitel einordnen lassen, die aber nicht unerwähnt bleiben sollen. Diese Arbeiten lassen sich am besten in das folgende Schema einordnen:

1. Virtuelles Screening
Es handelt sich dabei um die Entwicklung eines Webinterfaces, mit dem es möglich ist, die Selektivität eines Substrates für das Enzym Cyclooxygenase-II (COX-II) zu berechnen.
Des Weiteren wurde in der Zusammenarbeit mit Prof. Werner ein virtuelles Screening von etwas mehr als 126.000 Substanzen durchgeführt.

2. Verwaltung von Publikationen
Für die Verwaltung der Publikationen unserer Abteilung im DKFZ wurde ein Content-Management-System erstellt.

3. Automatische Analyse von Massenspektren
In Zusammenarbeit mit Prof. Lehmann entstand das Programm *AUTOMASCOT*, das es gestattet, automatisch Peptid-Sequenzen zu ermitteln und in einer Datenbank zu speichern.

9.1.1 AUTODOCK

Als Ausgangsbasis für die beiden Docking-Projekte diente das Programm *AUTODOCK* 3 des Scripps-Instituts. Autodock 3.0 ist eine Sammlung der Programme *AUTODOCK*, *AUTOGRID* und *AUTOTORS*[1], die im Jahre 1990 von David Goodsell entwickelt worden sind, und anschließend wurden sie von Garrett Morris bis zum heutigen Stand weiterentwickelt[137, 138]. Der Zweck dieser Programme ist es, die besten Konformationen von einem flexiblen Liganden, der an ein Makromolekül, z. B. ein Enzym oder einen DNA-Abschnitt, gebunden ist, zu berechnen.
Mit Hilfe des Programms *AUTOTORS* ist es möglich, bis zu acht verschiedene Torsionswinkel in dem Liganden zu bestimmen. Um diese Winkel kann bei der Suche nach den besten Dockingpositionen als weiterer Parameter zufällig gedreht werden. Die Kraftfeld-Parameter bestehen aus einer Untermenge des AMBER-Kraftfeldes [139].
Im nächsten Schritt erfolgt die Berechnung eines Würfels mit Rasterpunkten (Grids) durch das Programm *AUTOGRID*. Dazu wird als erstes festgestellt, welche Atomtypen im Makromolekül vorhanden sind. Für jeden Atomtyp wird dann ein Grid um das statische Makromolekül berechnet. Dabei sind verschiedene Parameter einstellbar. So können der Abstand der Gridpunkte und auch die Anzahl der Punkte bestimmt werden. Mit diesen Einstellungen ist es möglich, die Größe und Feinheit der Box zu bestimmen. Bei den vorliegenden Berechnungen wurden 61 Gridpunkte

[1] http://www.scripps.edu/pub/olson-web/doc/autodock/

mit einem Abstand von 0,375Å pro Kante gewählt. Daraus ergab sich ein Würfel mit einer Kantenlänge von 22,875Å, der insgesamt 226981 Rasterpunkte enthält.

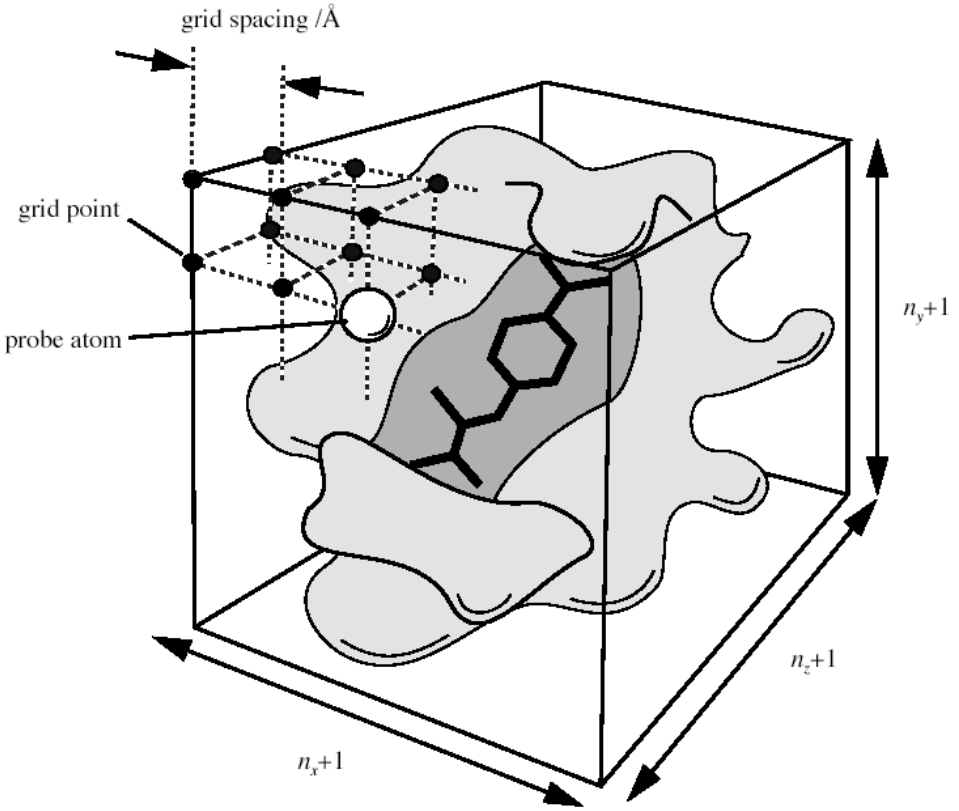


Abbildung 94: Schematische Darstellung des mit Autogrid um das Makromolekül berechneten Würfels

Für jeden Rasterpunkt werden nun die van der Waals Energien mit Hilfe der folgenden Formel mittels eines Lennard-Jones 6-12 Potentials berechnet:

$$E_{vdw}\left(r_{ij}\right) = \sum_{i<j, r_{ij}<r_{cutoff}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right]$$

wobei r_{ij} der Abstand zwischen zwei sich beeinflussenden Atomen i und j ist und A_{ij} und B_{ij} Konstanten sind, die aus den Potentialkurven der jeweiligen Atome berechnet werden.
Zusätzlich berechnet das Programm *AUTOGRID* ein Grid mit elektrostatischen Potentialen, die den Coulombschen Wechselwirkungen zwischen dem Makromolekül und einem Proben-Atom (mit einer Elementarladung e) entspricht. Zu diesem Zweck wurden mit Hilfe des Programms Insight die Partialladungen für jedes Atom im Makromolekül berechnet.
Nachdem auf diese Art und Weise die Eingabedateien erzeugt worden sind, kann mit Hilfe des Programms *AUTODOCK* durch ein Monte Carlo Simulated Annealing die beste Docking-Position in dem Grid gefunden werden. Bei dem Dockingvorgang wird nun der Ligand im Grid durch zufällige Änderungen seiner Orientierung, Position und Konformation bewegt. Mit Hilfe der implementierten Bewertungsfunktion wird nun der neu ermittelte Score mit dem vorhergehenden verglichen. Nach Beendigung des

Annealings ist auf diese Art und Weise die beste Konformation von Ligand und Makromolekül gefunden worden. Dieser Vorgang wurde zehnmal wiederholt und anschließend die berechneten Energien gemittelt. Je niedriger diese berechnete Energie ist, umso besser dockt der Ligand an das Makromolekül. Außerdem wurden die Ergebnisse visuell auf Plausibilität überprüft, damit es nicht aus Versehen zu einem Docking abseits der Bindetasche gekommen ist.
Das Programm ist für den akademischen Bereich kostenlos verfügbar und stellt nur geringe Anforderungen an die Hardware des verwendeten Rechners. So wurden die meisten Berechnungen auf einem Computer mit einem Pentium III Prozessor ausgeführt, der mit 450MHz getaktet ist. Die Ergebnisse sind in den folgenden beiden Abschnitten beschrieben.

9.2 Entwicklung einer webbasierten Anwendung zur Bestimmung der COX-II - Selektivität eines Substrates

Die Behandlung von rheumatischen und durch Entzündungen des Gewebes begründete Krankheiten findet eine erste Erwähnung im Corpus Hippocratum im dritten Jahrhundert vor Christus. Seit dieser Zeit wurden diese Krankheiten immer weiter untersucht und die Behandlung verbessert. So wird schon sehr lange ein Extrakt der Weidenrinde eingesetzt, der reich an Salicylaten, wie Salicylsäure und Gentisinsäure ist. Eine chemische Abwandlung dieser Salicylate ist die Acetylsalicylsäure, besser bekannt als Aspirin®, die seit mehr als hundert Jahren eingesetzt wird, um diese Krankheiten zu behandeln oder zumindest die Beschwerden zu lindern. Im Laufe der Zeit kamen immer mehr Substanzen dazu, so wurde im Jahre 1950 das Diclofenac auf den Markt gebracht, das bis heute als Standard bei der Behandlung von Rheuma und rheumatoider Arthritis eingesetzt wird. Leider führen viele dieser Substanzen, was für Aspirin schon im Jahre 1938 festgestellt wurde[140], zu starken Nebenwirkungen wie Blutungen im Magen-Darmtrakt und anderen gastrointestinalen Beschwerden, die teilweise so schwerwiegend sind, dass Patienten stationär behandelt werden müssen.
Lange Zeit konnte man sich diese Nebenwirkungen nicht erklären. Im Jahre 1971 wurde jedoch durch Vane et al.[141] der Wirkmechanismus dieser aspirin-ähnlichen Substanzen aufgeklärt. Dabei wurde festgestellt, dass die COX an der Entstehung des Prostaglandin E_1 beteiligt ist, das unter anderem für den Schutz der Magenschleimhaut verantwortlich ist. Werden nun über einen längeren Zeitraum NSAR genommen, kann es zu einer Degeneration der schützenden Schichten kommen und durch Einwirkung der Magensäure kann es ungehindert zu einem Selbstverdau des Magens kommen.
1991 wurde fast zeitgleich durch O'Banion et al. [142] und durch Kujubu et al.[143] ein durch Entzündungen induzierbares Isoenzym der Cyclooxygenase entdeckt. Diese COX-II wird im Gegensatz zur COX-I, die zu einem gewissen Level in allen Körpergeweben vorkommt durch bei Entzündungen gebildeten Mediatoren verstärkt gebildet.
Ein erster Versuch die COX-II gezielt zu hemmen wurde mit dem 10fach selektiveren Hemmstoff Meloxicam (Mobec®) durchgeführt. Es zeigte sich, dass die Substanz noch nicht selektiv genug für die COX-II ist, da die Substanz bei therapeutischen Dosen die COX-I immer noch zu 80 Prozent hemmt. Bei den ersten beiden zugelassenen selektiven COX-II-Hemmstoffen Refecoxib (Vioxx®) und Celecoxib (Celebrexx®), die eine bis zu 1000fach selektivere Hemmung besitzen, zeigte sich eine Verminderung der unerwünschten Wirkungen. So zeigten erste klinische

Studien mit diesen beiden Hemmstoffen, dass sich das Niveau der gastrointestinalen Nebenwirkungen auf Placebo-Niveau vermindern lässt[144].

9.2.1 Typischer Ablauf der Entzündungen

Die beiden Isoenzyme der COX spielen eine große Rolle in der Metabolisierung der Arachidonsäure. Eine Hemmung der COX ist wichtig bei der Behandlung von Krankheiten mit entzündlichen Prozessen wie Rheuma und Arthritis.

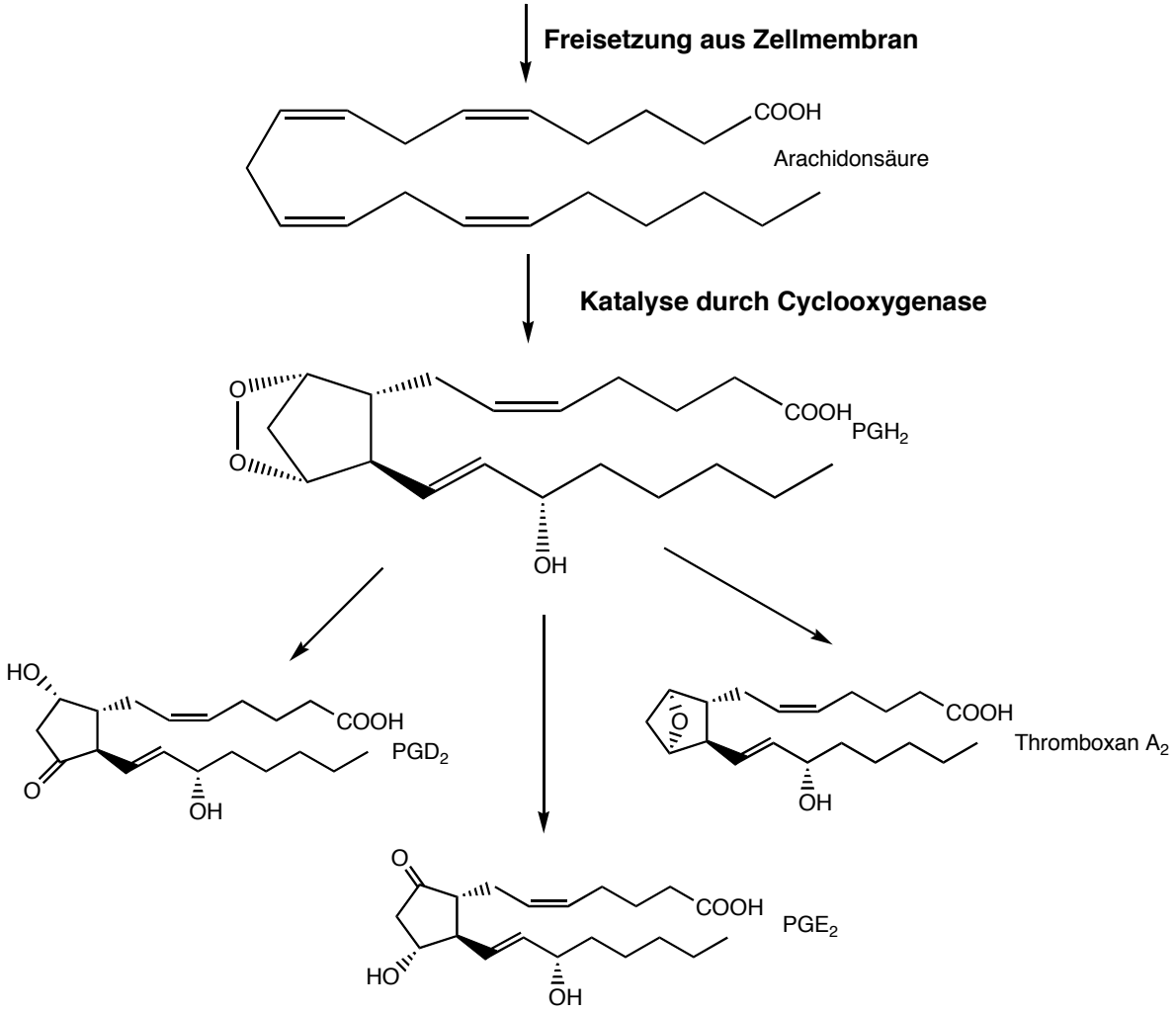


Abbildung 95: Metabolisierung von Arachidonsäure in der Zelle

9.2.2 Aufgabenstellung

Bislang geschieht die Testung meistens in in-vitro Testsystemen[145], die aber nur eine begrenzte Anzahl von Substanzen testen können und auch starke Schwankungen im Bereich der gemessenen Werte zeigen. Eine der größten Einschränkungen bei diesen Tests besteht aber in der Tatsache, dass die zu testende Substanz real vorhanden sein muss. Entweder muss die Substanz synthetisiert oder aber als Naturstoff extrahiert und in eines der Testsysteme eingesetzt werden. Dieser Umstand führt dazu, dass eine große Anzahl der in jedem Jahr neu synthetisierten und auch der neu isolierten Naturstoffe nicht auf Ihre Wirksamkeit bei rheumatischen Erkrankungen bzw. ihre Hemmwirkung der beiden Isoenzyme der COX untersucht werden können. Erste Versuche ein automatisiertes Docking mit Liganden durchzuführen verliefen sehr erfolgversprechend[146]. Dieses

Testsystem beruht unter anderem auch auf dem Programmpaket Autodock, hat aber den Nachteil, dass es nicht allgemein zugänglich ist. Es muss erst vom Benutzer installiert werden, und es verlangt daher einige Vorkenntnisse im Umgang mit Soft- und Hardware. Dieses führte zu der Überlegung ein virtuelles Testsystem aufzubauen, dass in der Lage ist, relativ schnell und einfach eine Vortestung auf eine COX-Hemmung und auf eine mögliche Selektivität dieses Stoffes für die COX-II zu ermöglichen. Außerdem sollte es sehr einfach über das Internet erreichbar sein.

9.2.3 Methodik

Die Suche nach Strukturen, die in der Brookhaven Protein Datenbank gespeichert sind, ergab für den Begriff ‚Cyclooxygenase' 19 verschiedene Einträge. Damit man vergleichbare Ergebnisse bezüglich der berechneten Werte für die Hemmwirkung erhält, wurden zwei Proteine ausgewählt, die zusammen mit einem Inhibitor kristallisiert worden sind. Für die COX-I wurde der Eintrag ‚1EQH'[147], gewählt, der einen Komplex darstellt, in dem das Enzym durch den Arzneistoff Flurbiprofen gehemmt ist. Flurbiprofen ist ein nichtselektiver Hemmstoff, der beide Isoenzyme ungefähr gleichmäßig hemmt.

Abbildung 96: Grafische Darstellung der COX-I[103]

Für die Berechnungen mit der COX-II wurde der Eintrag ‚1CX2'[148] genommen, bei dem das Enzym mit dem selektiven Hemmstoff SC-558 einen Komplex bildet.

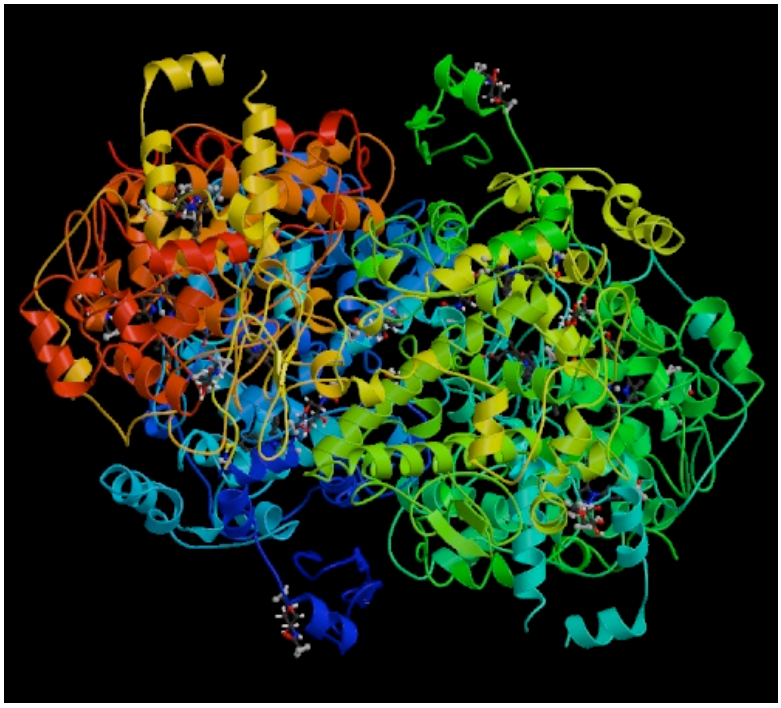


Abbildung 97: Grafische Darstellung der COX-2[103]

Bei beiden Einträgen wurde darauf Wert gelegt, dass es sich um reversible Hemmungen handelt, und dass nicht wie bei der Hemmung durch die Acetylsalicylsäure eine Suicid-Inhibierung an Position 534 des Enzyms stattfindet[149-151]. Durch die Verwendung des selektiven Hemmstoffes SC-558 sind auch die strukturellen Unterschiede bei der Hemmung der COX-II berücksichtigt[152], so dass bei den Berechnungen mit AutoGrid das Grid so berechnet werden kann, dass mögliche selektive Hemmstoffe in die entsprechend größere Seitentasche der COX-II gelangen können.
Nun wurden beide Strukturen in der gleichen Weise bearbeit: Als erstes wurde mit Hilfe des Programms Insight für beide Strukturen die Partialladungen der Atome berechnet. Dann wurde mit dem Programm AutoGrid, wie oben beschrieben, das Grid berechnet, in dem der Ligand sich später bewegen kann.

9.2.4 Berechnung der Selektivität

Zur Berechnung der Selektivität wird die Hemmung des Substrates für jedes der beiden Isoenzyme berechnet. Dazu wird die berechnete Energie der Referenzsubstanz (dabei handelt es sich um den mit dem Isoenzym kristallisierten Inhibitor), ins Verhältnis zum berechneten Ergebnis des untersuchten Substrats gesetzt. Der Quotient aus der berechneten Hemmung von COX-I und COX-II liefert nun ein Maß für die Selektivität der Verbindung:

$$\text{Selektivität} = \frac{\text{Energie}_{\text{SubstratCox-I}} - \text{Energie}_{\text{ReferenzCox-I}}}{\text{Energie}_{\text{SubstratCox-II}} - \text{Energie}_{\text{ReferenzCox-II}}}$$

9.2.5 Das Webinterface

Eine der wichtigsten Anforderungen bestand darin, dass das Testsystem von überall in der Welt zugänglich sein sollte. Dies ist ohne größere Probleme über ein Webinterface erreichbar, das in Abbildung 98 sichtbar ist. Dabei wurden so wenige

Vorkenntnisse wie eben nötig vorausgesetzt. So ist es nicht nötig irgendwelche Parameter zu setzen, sondern jedes Docking wird mit standardisierten Parametern durchgeführt. Natürlich muss der Benutzer wissen, wie er eine Datei mit den Strukturdaten des Liganden im mol2-Format erzeugt. Dies ist aber mit allen gängigen Programmen wie *CHEMOFFICE*® und *ISISDRAW*® möglich. Eine durchschnittliche Berechnung benötigt bei der jetzigen Hardware im Moment 20 Minuten. Wird das Testsystem von mehreren Personen gleichzeitig benutzt, kann sich die Berechnungszeit jedoch bedeutend verlängern. Er muss daher als zweiten Parameter seine Email-Adresse angeben.

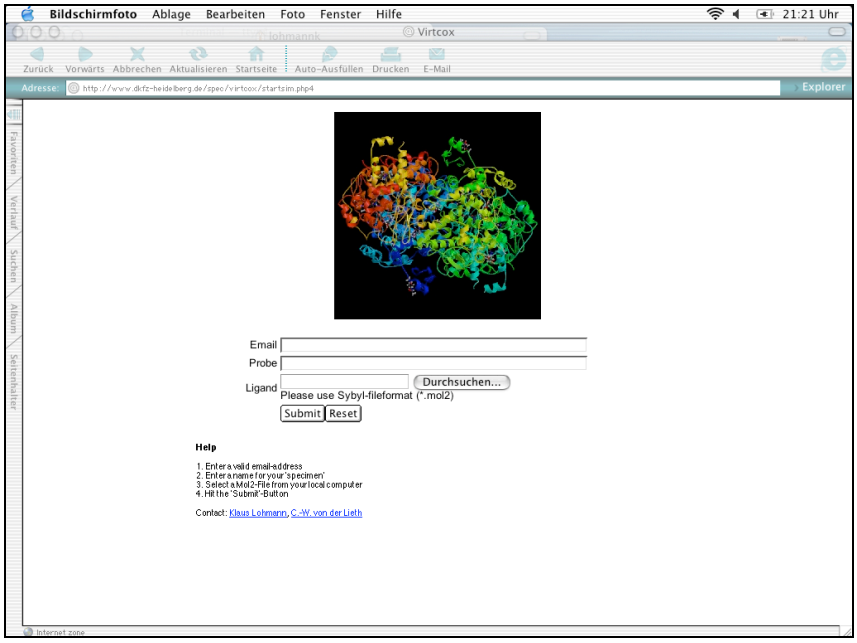


Abbildung 98: Das Webinterface zur Eingabe der Liganden

Um seine Berechnungen auch im Nachhinein noch unterscheiden zu können, hat der User die Möglichkeit den Namen der Substanz einzugeben. Durch Anklicken des <Submit>-Button wird die Berechnung gestartet. Nachdem die Berechnungen beendet sind, erhält man eine Email in der folgenden Form:

```
Dear Sirs,

Your calculations for Refecoxib have been finished!!
The substrate is 10.5000fold more selective for COX-II as COX-I

Sincerely
zshubble
```

Mit Hilfe dieses Interfaces ist es nun sehr einfach möglich, seine Substanzen auf eine mögliche selektive Hemmung der COX-II zu testen, sofern man über einen Computer mit Internet-Zugang verfügt.

9.2.6 Ergebnis und Diskussion

Mit Hilfe des Webinterfaces wurden siebzehn auf dem Markt befindliche NSAR getestet:

Inhibitor	COX-II-Selektivität	Berechnete Selektivität (>3 bedeutet COX-II Selektivität)
Tiaprofen	Non-selective	0,63
Phenazon	Non-selective	0,79
Ibuprofen	Non-selective	0,81
Ketoprofen	Non-selective	0,83
Acetylsalicylic acid	Non-selective	0,93
Salicylsäure	Non-selective	1,00
Diclofenac	Non-selective	1,13
Flurbiprofen	Non-selective	1,23
Propyphenazon	Non-selective	1,26
Tenoxicam	Non-selective	1,28
Piroxicam	Non-selective	1,36
Naproxen	Non-selective	1,48
Diclofenac	Non-selective	1,58
Meloxicam	Non-selective	2,76
Celecoxib	Selective	3,66
Rofecoxib	Selective	10,50
S58	Selective	11,31

Tabelle 21: Ergebnisse der Selektivitätsberechnungen

Dabei zeigte sich, dass die Selektivität der im Moment auf dem Markt befindlichen Arzneistoffe sehr gut erkannt worden ist. Leider standen nicht mehr in-vitro Ergebnisse zur Verfügung, so dass weitere Berechnungen nicht stattfinden konnten. Bei allen untersuchten Substanzen ist die Selektivität der Inhibitoren richtig erkannt worden, so dass die Methode für ein schnelles Screening von Substanzen verwendet werden kann. Besonders trifft dieses zu, wenn die Subtanzen ansonsten in keinem Testsystem untersucht würden. Sie eignet sich daher besonders für Forscher, die an der Synthese neuer Arzneistoffe arbeiten oder für pharmazeutische Biologen[153, 154], die im Bereich der Extraktion von Naturstoffen aus Pflanzen oder anderen Organismen tätig sind. Natürlich handelt es sich dabei nur um eine Vortestung, aber es ist damit gut möglich einen Hinweis zu erhalten, ob eine Substanz für weitere Untersuchungen in Frage kommt oder nicht.

9.3 High Throughput Screening des Proteins YY-1

Das Protein Ying-Yang 1 (YY-1) ist in der Lage sich an die Promotorregion des Adeno assozierten Virus (AAV)[155] und an die Promotorregion der Cytochrom C Oxidase Untereinheit Vb zu binden[156]. Das Protein YY-1 ist zwingend notwendig für den Start der Transkription durch die RNA-Polymerase, so dass ein Blockieren der Bindestelle dieses Verhindern würde[156, 157].

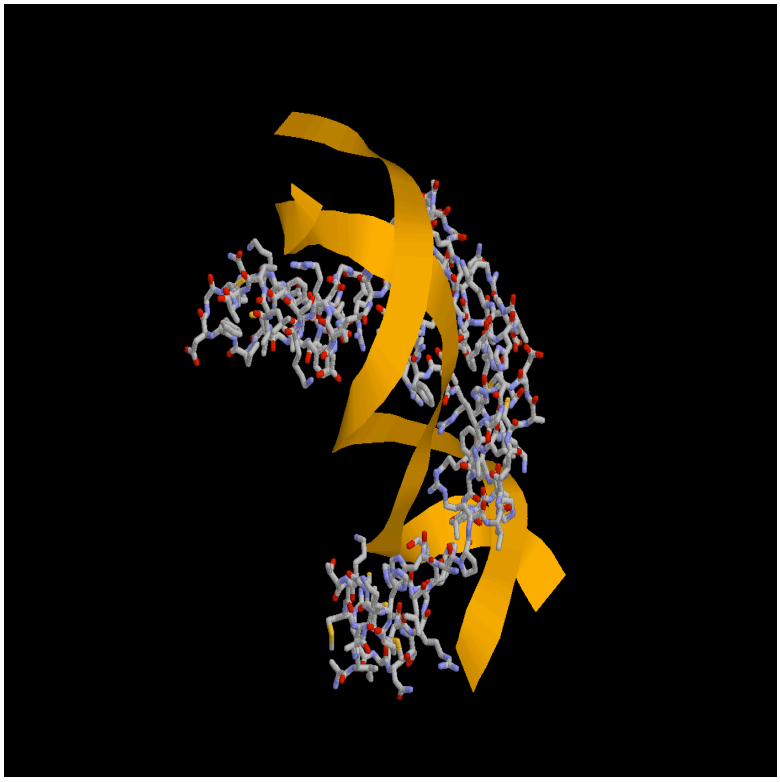


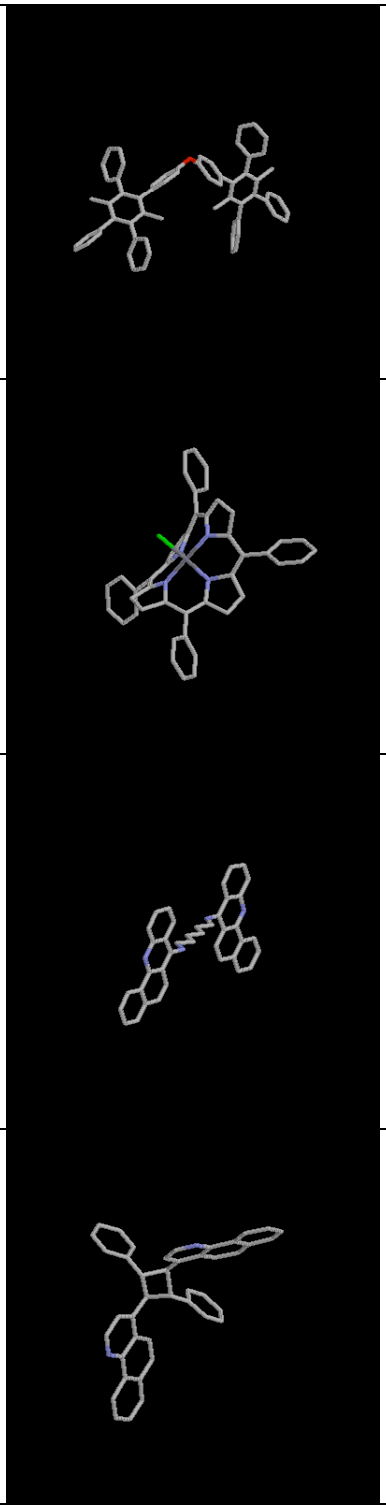
Abbildung 99: Position des Ying-Yang-Proteins um den DNA-Strang

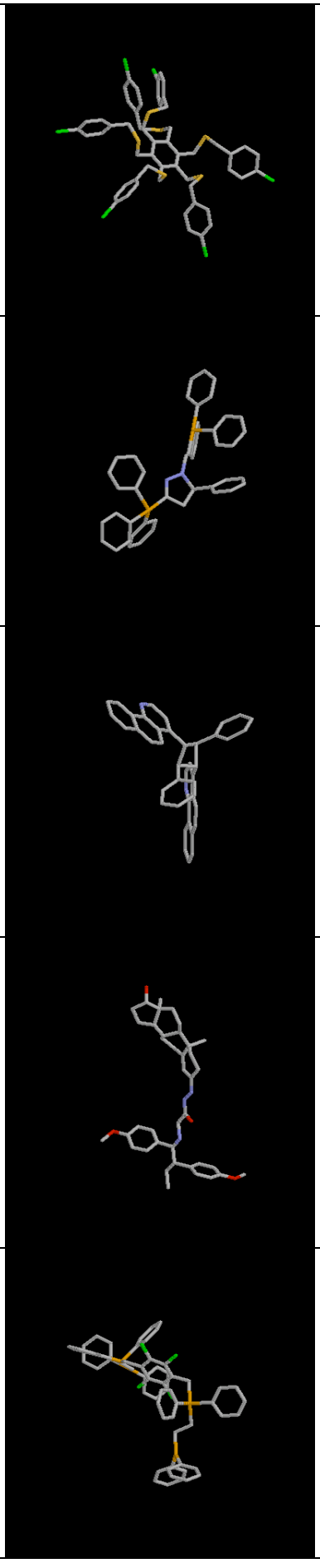
9.3.1 Methodik

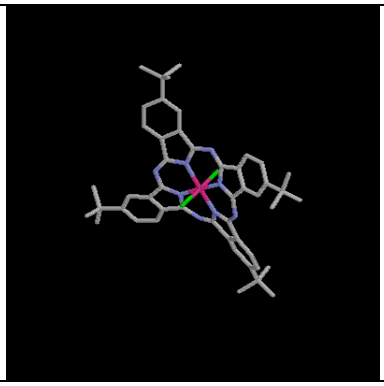
Das Screening basiert auf der Annahme, dass sich ein Molekül in den Bereich zwischen Protein- und DNA-Strang setzt, so dass die Funktion des Proteins als Initiatorelement verhindert wird. Für die Berechnungen wurde der Eintrag ‚1UBD'[158] aus der *PDB* genommen und mit Hilfe des Programms *INSIGHT* bearbeitet. Dabei wurden aus der Datei die Atome des DNA-Strangs entfernt, so dass nur noch der Protein-Strang für das Docking zur Verfügung stand. Für das Docking wurde nun der Bereich, in dem der Ligand sich bewegen kann, so gewählt, dass er sich wie in Abbildung 100 in dem Bereich befand, wo sonst die Nukleotid-Sequenz umschlossen wird.
Die Berechnungen wurden aus Geschwindigkeitsgründen auf dem Linux-Cluster unserer Abteilung durchgeführt. Es handelt sich dabei um einen Cluster mit 9 Knoten und insgesamt 18 Athlon-Prozessoren.
Als Liganden wurden die 3D-Strukturen genommen, die sich in der Datenbank NCIOpenmol befinden und durch das Programm Corina[159] erzeugt worden sind. Dabei wurden 126.056 verschiedene Strukturen gescreent. Um eine bessere Verwendung der 3D-Strukturen bei weiteren Screenings zu ermöglichen, wurden die in der NCIOpenmol enthaltenen Daten in eine MySQL-Datenbank übertragen, da so einfacher ein Zugriff erfolgen kann. Man hatte so auch die Möglichkeit fehlerhafte

Strukturen, die sich leider unter den Daten befanden, aus dem Datenbestand zu löschen, ohne dass man Routinen entwickeln musste, die das spezielle Dateiformat der ursprünglichen[159] Datenbank bearbeiten können.

9.3.2 Ergebnis

Rang		Substanz	gedockte Energie
1		98393	-4,0755
2		84770	-3,7395
3		109291	-3,705
4		108551	-3,6695

5		109622	-3,6635
6		92762	-3,6565
7		108552	-3,642
8		109660	-3,5495
9		100037	-3,5385

10		115578	-3,5045

Die Ergebnisse des Screenings lassen sich unter folgender URL im Internet ansehen: http://www.dkfz-heidelberg.de/spec/projekte/yy1/result/.

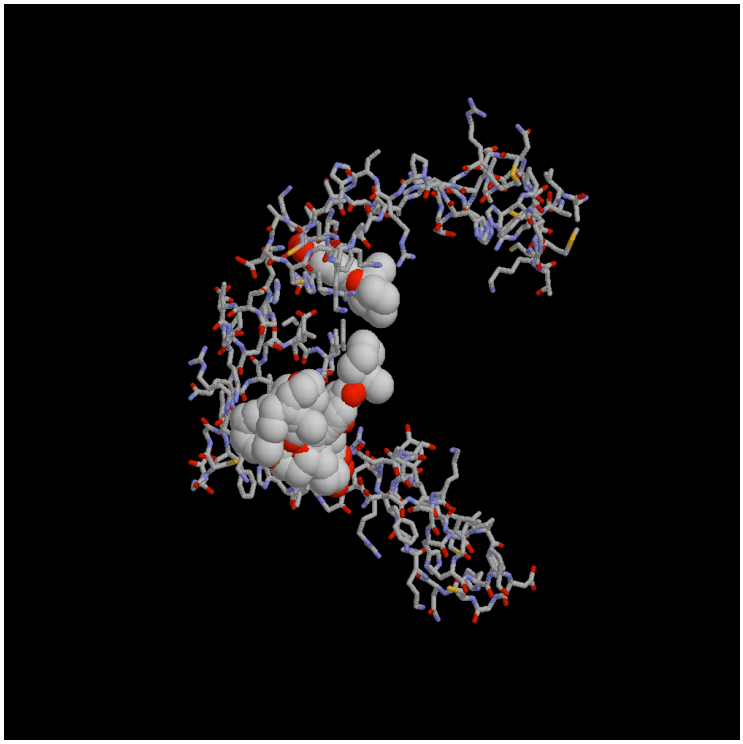


Abbildung 100: Ying-Yang-Proteins mit einem Inhibitor

Auf den ersten Blick ist zu erkennen, dass die Strukturen die besten Komplexe bilden, die sich gut um die Helix des Proteins legen können, wie dies in Abbildung 100 zu sehen ist. Leider wurde eine Testung der Substanzen in-vitro durch fehlende Testkapazitäten des beteiligten Projektpartners verhindert.

9.4 Entwicklung eines Content-Managementsystem zur Verwaltung von Publikationslisten

Für jeden wissenschaftlich arbeitenden Arbeitskreis ist es wichtig, seine geleisteten Arbeiten einem interessierten Publikum auf seiner Instituts-Homepage zu präsentieren. Außerdem müssen mit unterschiedlichen Kriterien Listen generiert werden, um z. B. die Publikationen zu einen bestimmten Projekt anzuzeigen. Da wie oben beschrieben die Techniken zur Extraktion von Daten aus der *PUBMED* entwickelt worden sind, bot es sich an, dies auch für die Entwicklung eines kleinen CMS zu nutzen.

Das System sollte den folgenden Ansprüchen genügen:

1. Darstellung der Publikationsliste als Webseite mit verschiedenen Suchmöglichkeiten.
2. Einfache passwort-geschützte Verwaltung der einzelnen Artikel und Kongressbeiträge
3. Extraktion von Daten aus der *PUBMED*

Bislang sind keine Entwicklungen gemacht worden, die diese Dinge in Echtzeit verwalten und auch für kleinere Arbeitsgruppen zur Verfügung stehen. Es sollte daher untersucht werden, inwieweit man diese Dinge mit frei zugänglichen Datenbanken und freier Software realisieren kann.
Natürlich hätte man dafür ein frei verfügbares Content Management System verwenden können. Hier wäre man mit der Einrichtung und Verwaltung länger beschäftigt gewesen, als es für diese relativ kleine Aufgabe nötig ist. Es war daher einfacher, zwei verschiedene Skripte zu entwickeln: Skript eins erledigt die Anzeige der Publikationen auf der Homepage und das zweite übernimmt die Verwaltungsaufgaben.

9.4.1 Technische Umsetzung

Wie auch bei den anderen Entwicklungen dieser Arbeit wurde hierbei auf die Datenbanksoftware MySQL und als Hypertext Preprozessor PHP zurückgegriffen. Die Daten zu den einzelnen Veröffentlichungen wurden mit Hilfe der von der NCBI zur Verfügung gestellten Schnittstellen extrahiert.

9.4.2 Anzeige der Publikationsliste

Die Anzeige der Publikationsliste besteht aus zwei Teilen. Im oberen Teil des Dokuments besteht die Möglichkeit, die verschiedenen Suchkriterien einzustellen. Dabei kann der Zeitraum der gesuchten Publikationen ausgewählt werden, aber auch nach Schlagwörtern oder Autoren gesucht werden. Zusätzlich kann die Liste noch nach Themengebieten und nach der Projektzugehörigkeit eingeschränkt werden.

Abbildung 101: Webformular zur Eingabe der Suchkriterien

Sucht man beispielsweise nach allen Publikationen, die als Autor oder im Titel das Wort ‚Lohmann' besitzen und in den Jahren zwischen 2002 und 2003 veröffentlicht worden sind, so erhält man die folgende Liste:

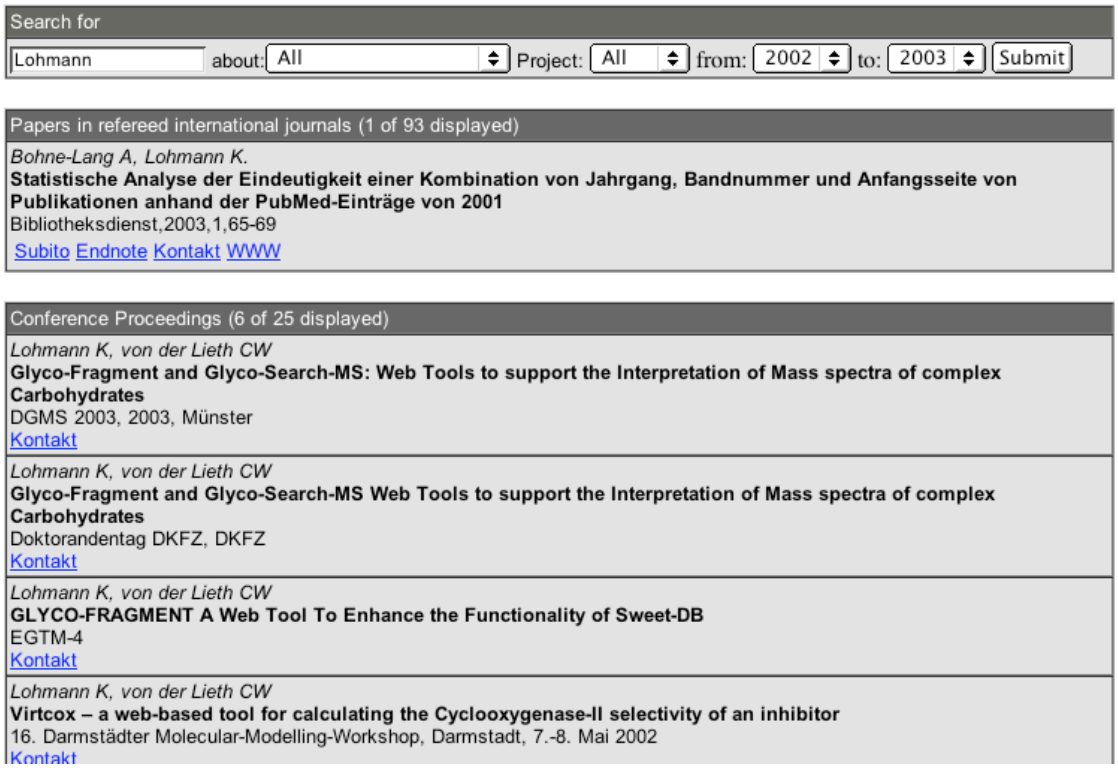


Abbildung 102: Beispiel für eine mögliche Ergebnisliste

Die Eingabe der Suchkriterien wurde bevorzugt mit Pulldown-Menüs realisiert, damit es für den Benutzer möglichst einfach ist, diese einzugeben. Es wird so vermieden, dass Suchbegriffe eingegeben werden, die zu leeren Ergebnislisten führen. Die Pulldown-Menüs werden ebenfalls dynamisch erzeugt und passen sich den in der Datenbank gespeicherten Daten an. So wird z.B. automatisch als Suchkriterium das Jahr 2004 hinzugefügt, sobald eine dementsprechende Veröffentlichung sich in der Datenbank befindet.
Sehr großen Wert wurde dabei auf die Verlinkung der Inhalte und einfache Weitergabe der Daten gelegt. So befinden sich bei Artikeln aus Zeitschriften unterhalb der bibliographischen Informationen Links. Mit denen kann sich der Benutzer an den korrespondierenden Autor wenden, oder er kann zu einer zu dem Artikel gehörenden Website surfen. Mit Hilfe dieser Links ist es aber auch möglich, die Daten in ein Literaturprogramm wie *ENDNOTE* oder *REFERENCE MANAGER* zu übernehmen. Damit wird dem Benutzer unnötige Tipparbeit abgenommen.

9.4.3 Verwaltung der Publikationen

Dieser Bereich ist der Teil, der nicht für alle zugänglich sein darf. Daher ist das Skript mit einem Passwortschutz versehen:

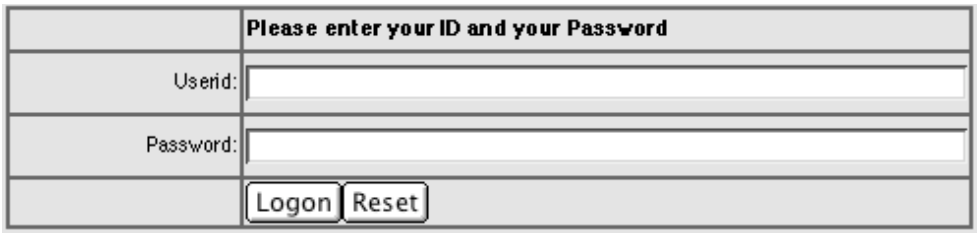


Abbildung 103: Logon für die Literaturverwaltung

Damit wird verhindert, dass Personen, die zufällig Zugriff auf diese Webseiten erhalten, Daten verändern oder löschen. Es können für einzelne Benutzer Passwörter vergeben werden, aber auch für eine Abteilung kann ein einziges Passwort genutzt werden.
Die Verwaltung der Publikationen sollte so einfach wie möglich gestaltet sein. Natürlich musste es zum einen möglich sein, sämtliche Datenfelder zu verändern, zum andern aber auch Publikationen wieder zu entfernen, die aus Versehen oder völlig falsch eingegeben worden sind.

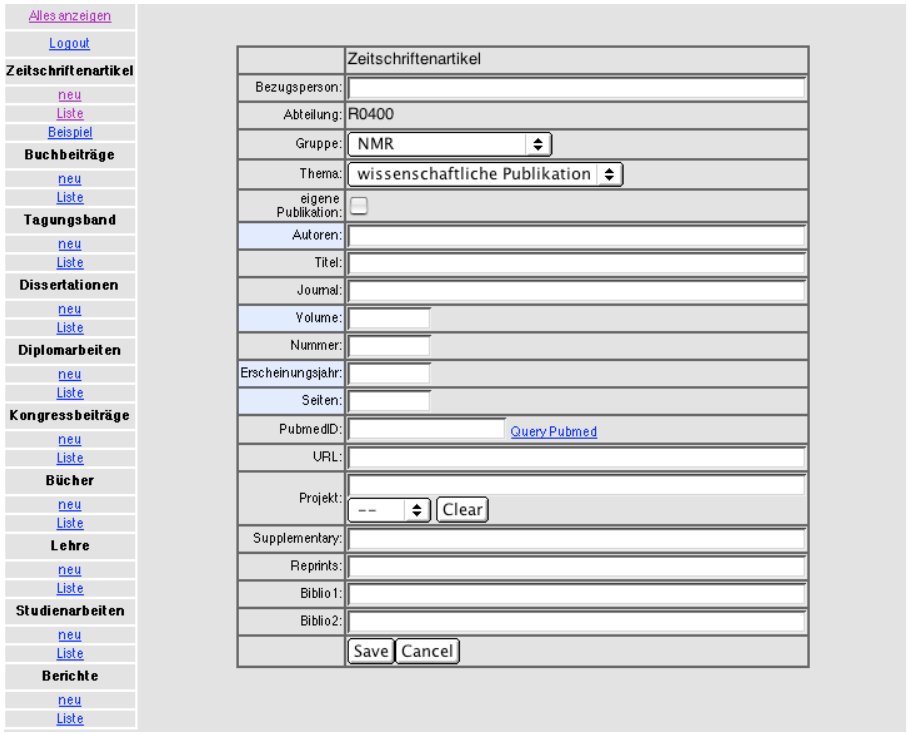


Abbildung 104: Webformular zur Eingabe der bibliographischen Daten

Dazu befindet sich auf der linken Seite ein Menü mit dem es möglich ist, sich die zu den einzelnen Veröffentlichungen nötigen Formulare in den rechten Teil des Bildschirms zu laden. Bei der Neueingabe erhält man sofort ein leeres Formular angezeigt. Hier hat man bei Artikeln in Zeitschriften die Möglichkeit, sich die Daten aus der *PUBMED* herunter zu laden. Dazu können in die Felder Autoren, Erscheinungsjahr, Volume und Seiten, Werte eingetragen werden, mit denen dann eine Anfrage an die *PUBMED*-Datenbank gemacht wird. Aus den Ergebnissen kann dann der entsprechende Artikel ausgewählt werden, und die Daten werden automatisch in die entsprechenden Felder eingetragen. Dieses erleichtert die Eingabe, und es reduziert stark die Tippfehler. Damit ist auch gewährleistet, dass z.B. die Jahreszahlen stimmen, und damit auch die Pulldown-Menüs für die Suchkriterien richtig erzeugt werden.

9.4.4 Ergebnis

Mit diesem kleinen Content-Management-System konnte gezeigt werden, dass es ohne größere Probleme möglich ist, die Literatur einer Arbeitsgruppe zu verwalten und praktisch in Echtzeit auf der Website der Arbeitsgruppe anzuzeigen. Die Daten sind vor Manipulationen geschützt, und es besteht für den Nutzer im Internet die Möglichkeit, Daten in sein Literaturverwaltungsprogramm zu übernehmen. Die Links der Anzeige gestatten eine einfache Kommunikation mit den Autoren und erlauben den Zugriff auf weitere Informationen.

9.5 *AUTOMASCOT*

Die Sequenzierung von Proteinen unter Zuhilfenahme der von Matrixscience entwickelten Software *MASCOT* gestaltet sich durch das mitgelieferte Webinterface sehr leicht. Sämtliche Parameter können ohne größere Probleme eingestellt werden und finden bei der Auswertung Berücksichtigung. Leider müssen alle Auswertungen von Hand durchgeführt werden, da es keine Möglichkeit gibt, das Absenden automatisch vorzunehmen. Bei dieser Durchführung gestaltet sich die Auswertung der zurückerhaltenen Ergebnisse besonders langwierig. Bei der massenspektrometrischen Analyse von Proteinen fallen inzwischen eine große Anzahl von MS/MS-Spektren an, bei denen häufig die Zeit für die Auswertung fehlt. Dabei gehen aber wichtige Informationen verloren, was durch den Einsatz geeigneter bioinformatorischer Methoden verhindert werden könnte.

9.5.1 Anforderungen

Im Rahmen dieses Projektes sollte untersucht werden, ob es möglich ist, aus routinemäßig aufgenommenen Massenspektren von Proteinen eine Datenbank zu erstellen, in der die in den Spektren enthaltenen Sequenzinformationen gespeichert werden. Dabei sollten die wichtigsten Parameter, die *MASCOT* bietet, auch hier einstellbar sein. Da *MASCOT* im DKFZ nur in einer Webversion verfügbar ist, musste eine Schnittstelle entwickelt werden, die es gestattet, Anfragen automatisch zu erzeugen und die Antwort des *MASCOT*-Servers automatisch auszuwerten. Diese Option ist routinemäßig nicht vorhanden und musste daher ebenfalls nachträglich implementiert werden. Dies konnte sehr einfach mit Hilfe eines kleinen CGI-Skriptes erreicht werden, das über einen Aufruf des Webservers das gesamte von *MASCOT* erzeugte Ergebnis-File zurückliefert.
Der Ablauf der Auswertung stellt sich wie folgt dar: Mittels eines einfachen POST-Request werden sämtliche für die Berechnung notwendigen Informationen, wie das MS/MS-Spektrum und die eingestellten Parameter, an den *MASCOT*-Server übermittelt. Nachdem die Berechnungen abgeschlossen sind, erfolgt die Übertragung der kompletten Ergebnis-Datei durch einen GET-Request an das Programm *AUTOMASCOT*. Bei dieser Ergebnis-Datei handelt es sich um eine Textdatei, die sämtliche Informationen enthält, die benötigt werden um eine automatische Auswertung vornehmen zu können. So entspricht die Zeile ‚h1_q5=0,762.402405, -0.010645, 228,233, 4.00, SYPQLR, 26,00000000, 21.87, 1,000000201000000, 0, 0, 677.418000' einer Zuordnung der Sequenz ‚SYPQLR' zu dem Peak mit der Masse 762,402405. Die Informationen können nun nacheinander ausgewertet werden und in die Datenbank eingetragen werden. Dabei wird der ermittelte *MASCOT*-Score entsprechend der Vorgaben berücksichtigt. Wird ein Score zurückgeliefert, der kleiner als 80 ist, so wird das gesamte Ergebnis verworfen, da es auch von einer zufälligen Übereinstimmung der beiden verglichenen Spektren kommen kann.

9.5.2 Technische Umsetzung

Da es sich bei diesen Projekt um ein sehr komplexes Programm handelt, kam eine reine webbasierte Lösung nicht in Frage, und es wurde daher ein auf dem Cocoa-Framework basierendes Programm entwickelt, das die Interaktion mit dem Benutzer, aber auch die Kommunikation mit dem Server und der Datenbank vornimmt.

9.5.3 Einstellen der *MASCOT* Parameter

Ein sehr wichtiger Aspekt dieses Programms bestand darin, dass die von *MASCOT* zulässigen Parameter ebenfalls eingestellt werden können. Dieses wurde über die Entwicklung einer Preference-Klasse erreicht, die das Einstellen erlaubt. Dazu brauchte die Klasse, die es gestattet, die Preferences für das Programm *FINDYSERIES* einzustellen, nur um den Parameter ‚Data Directory' erweitert werden.

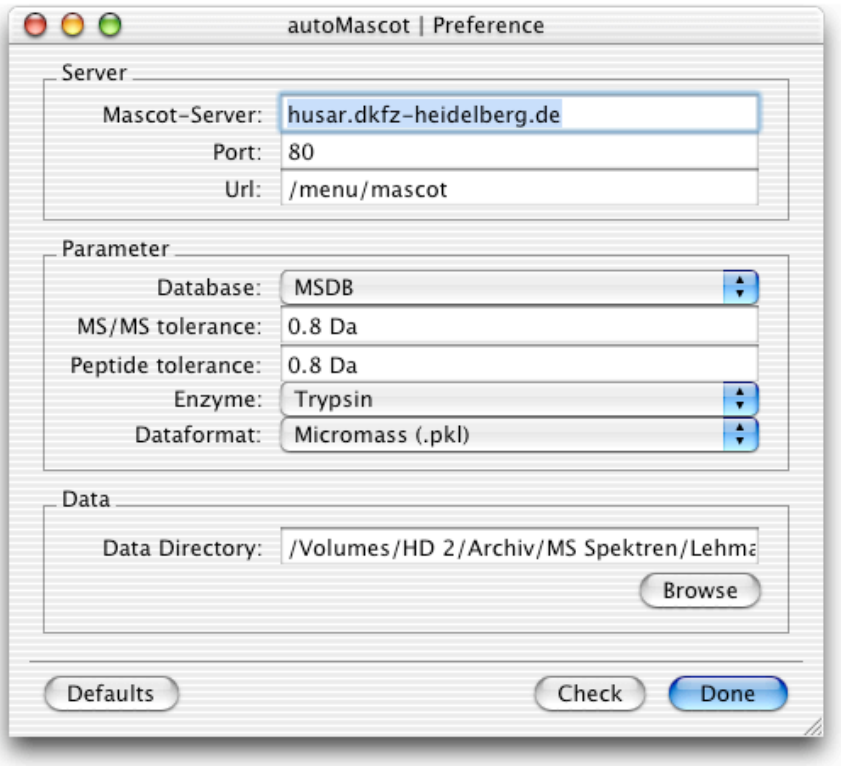


Abbildung 105: Dialog zum Einstellen der *MASCOT*-Parameter

Hier ist es möglich die Bedingungen des Proteinverdaus, das Datei-Format und das Verzeichnis in dem die Spektren abgelegt worden sind, einzustellen. Das Programm sendet nun nacheinander alle Spektren an den *MASCOT*-Server und wertet die Ergebnisse aus.

9.5.4 Darstellung der Ergebnisse

Die übermittelten Daten, wie Spektrum und Messparameter, werden zusammen mit den *MASCOT*-Ergebnissen in einer MySQL-Datenbank gespeichert. Dieses gestattet einen sehr einfachen und auch universellen Zugriff auf die Daten, da die Datenbank sowohl in PHP als auch in C durch eine entsprechende Schnittstelle angesprochen werden kann. Um einen standort- und betriebssystemunabhängigen Zugriff zu ermöglichen, wurde die Darstellung der Ergebnisse durch ein Webinterface gewählt. Der Benutzer erhält hier die Möglichkeit sich alle Ergebnisse, die automatisch gesammelt worden sind, in einer Liste darzustellen.

Search for: Annexin view all

Proteinname	Proteinmass	Mascot-Score	Mascot-Result	
Annexin I (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhibitory protein h1_q1=	38558.9400	818	view details	view MS/MS results
Annexin I (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhibitory protein h1_q1=	38558.9400	818	view details	view MS/MS results
Annexin II (Lipocortin II) (Calpactin I heavy chain) (Chromobindin 8) (P36) (Protein I) (Placental h1_q1=	38456.7200	123	view details	view MS/MS results
Annexin II (Lipocortin II) (Calpactin I heavy chain) (Chromobindin 8) (P36) (Protein I) (Placental h1_q1=	38448.7700	752	view details	view MS/MS results
Annexin II (Lipocortin II) (Calpactin I heavy chain) (Chromobindin 8) (P36) (Protein I) (Placental h1_q1=	38456.7200	305	view details	view MS/MS results
Annexin II (Lipocortin II) (Calpactin I heavy chain) (Chromobindin 8) (P36) (Protein I) (Placental h1_q1=	38448.7700	833	view details	view MS/MS results
Annexin II (Lipocortin II) (Calpactin I heavy chain) (Chromobindin 8) (P36) (Protein I) (Placental h1_q1=	38456.7200	123	view details	view MS/MS results
Annexin II (Lipocortin II) (Calpactin I heavy chain) (Chromobindin 8) (P36) (Protein I) (Placental h1_q1=	38448.7700	752	view details	view MS/MS results
Annexin II (Lipocortin II) (Calpactin I heavy chain) (Chromobindin 8) (P36) (Protein I) (Placental h1_q1=	38456.7200	305	view details	view MS/MS results
Annexin II (Lipocortin II) (Calpactin I heavy chain) (Chromobindin 8) (P36) (Protein I) (Placental h1_q1=	38448.7700	833	view details	view MS/MS results

Abbildung 106: Webinterface zur Darstellung der *AUTOMASCOT*-Ergebnisse

Der Benutzer hat die Möglichkeit durch das Eingeben eines Suchwortes die Liste einzuschränken und die Liste nach Proteinname, Proteinmasse und nach der Höhe des *MASCOT*-Scores sortieren zu lassen. Dieses gestattet eine sehr effiziente und einfache Möglichkeit, sich das richtige Ergebnis anzeigen zu lassen. Der ermittelte *MASCOT*-Score zeigt sehr gut an, ob die Messbedingungen optimal waren, so dass eine große Anzahl der möglichen Bruchstücke auch gemessen worden sind. Ausgehend von dieser Liste kann er sich weitere Details der einzelnen Ergebnisse anzeigen lassen. Es ist möglich, sich alle Informationen, die von *MASCOT* zurückgeliefert worden sind, im Rohformat darzustellen. Viel Interessanter für den Benutzer ist aber die Darstellung der gefunden Sequenzinformationen.

Sequenz	ion	measured mass	calculated mass	deltaM
SYPHLQK	404.7359	807.4562	807.4603	-0.0041
DLYDAGYK	440.7182	879.4208	879.4337	-0.0129
AYTNFDAER	543.2629	1084.5101	1085.4777	-0.9676
DALNIETAIK	544.2885	1086.5614	1086.5920	-0.0306
QDIAFAYQR	556.2740	1110.5323	1110.5458	-0.0135
TPAQYDASELK	611.7881	1221.5605	1221.5877	-0.0272
TNQELQEINR	622.8105	1243.6053	1243.6156	-0.0103
SLYYYIQQDTK	711.3350	1420.6544	1420.6874	-0.0330
SLYYYIQQDTK	711.3448	1420.6739	1420.6874	-0.0135
GYDEYTIYNILTNR	771.8931	1541.7706	1541.8412	-0.0706
LSLEGDHSTPPSAYGSVK	615.6323	1843.8735	1843.8951	-0.0216
AEDGSVIDYELIDQDAR	954.9608	1907.9060	1907.8748	0.0312

Abbildung 107: Detail-Ansicht eines Suchergebnisses

Hier sind alle im Spektrum zugeordneten Peaks enthalten. Die Darstellung der Liste ist nach aufsteigender Masse der Ionen sortiert. Danach werden die gemessenen und berechneten Massenwerte, sowie die Differenz der beiden Werte angezeigt. Dies gestattet es dem Benutzer sehr leicht festzustellen, wie gut seine verwendeten Messparameter sind, um unter Umständen die Messbedingungen zu optimieren.

9.5.5 Ergebnis

Mit der Entwicklung dieser Software konnte gezeigt werden, dass es ohne größere Schwierigkeiten möglich ist, Massenspektren routinemäßig auszuwerten. Der Zeitaufwand, der für eine manuelle Analyse benötigt wird, kann bis zu einer Stunde betragen. Dieses konnte durch diese Methode auf wenige Minuten reduziert werden. Mit diesen Methoden wurden innerhalb von etwa 38 Stunden 224 MS/MS-Spektren ausgewertet. Dabei lieferten 115 der 224 einen *MASCOT*-Score der höher als 80 war, und sie konnten daher ausgewertet werden. Dieses entspricht einer Quote von 51,3 Prozent. Dabei wurden 1104 Sequenzen einem MS-Spektrum zugeordnet.

10 Abbildungsverzeichnis

11 Dateiformate und Handbücher

In diesem Abschnitt sind die Datei-Formate und die XML-Container zum Austausch von Daten über das Internet wiedergegeben. Außerdem sind hier hier die zur Benutzung der Programme nötigen Handbücher abgelegt.

11.1 XML-Format zum Austausch von Massenspektren

```
  <?xml version="1.0" encoding="utf-8" ?>
  <!DOCTYPE SweetDBMassSpectra (View Source for full doctype...)>
- <SweetDBMassSpectra>
- <PubmedArticle>
  <PMID>11110090</PMID>
  <Year>2000</Year>
- <Journal>
  <JournalName>J Mass Spectrom</JournalName>
  <Volume />
  <Issue>35</Issue>
- <PubDate>
  <Year>2000</Year>
  </PubDate>
  </Journal>
  <ArticleTitle>Collision-induced fragmentation of underivatized N-linked
carbohydrates ionized by electrospray</ArticleTitle>
- <Pagination>
  <MedlinePgn>1178-90</MedlinePgn>
  </Pagination>
- <Abstract>
  <AbstractText>The electrospray mass spectra and collision-induced
fragmentation of neutral N-linked glycans obtained from glycoproteins were
examined with a Q-TOF mass spectrometer. The glycans were ionized most
effectively as adducts of alkali metals, with lithium providing the most
abundant signal and caesium the least. Singly charged ions generally gave
higher ion currents than doubly charged ions. Addition of formic acid could
be used to produce [M + H]+ ions, but these ions were always accompanied by
abundant cone-voltage fragments. The energy required for collision-induced
fragmentation was found to increase in a linear manner as a function of
mass with the [M + Na]+ ions requiring about four times as much energy as
the [M + H]+ ions for complete fragmentation of the molecular ions.
Fragmentation of the [M + H]+ ions gave predominantly B- and Y-type
glycosidic fragments whereas the [M + Na]+ and [M + Li]+ ions produced a
number of additional fragments including those derived from cross-ring
cleavages. Little fragmentation was observed from the [M + K]+ and [M +
Rb]+ ions and the only fragment to be observed from the [M + Cs]+ ion was
Cs+. The [M + Na]+ and [M + Li]+ ions from all the N-linked glycans gave
abundant fragments resulting from loss of the terminal GlcNAc moiety and
prominent, though weaker, ions as the result of 0,2A and 2,4A cross-ring
cleavages of this residue. Most other ions were the result of successive
additional losses of residues from the non-reducing terminus. This pattern
was particularly prominent with glycans containing several non-reducing
GlcNAc residues where successive losses of 203 u were observed. Many of the
ions in the low-mass range were products of several different fragmentation
routes but still provided structural information. Possibly of most
diagnostic importance was an ion formed by loss of 221 u (GlcNAc molecule)
from an ion that had lost the 3-antenna and the chitobiose core. This
latter ion, although coincident in mass with some other 'internal'
fragments, often provided additional information on the composition of the
antennae. Other ions defining antenna composition were weak cross-ring
fragments produced from the core branching mannose residue. Glycans
containing Gal-GlcNAc residues showed successive losses of this moiety,
particularly from the B-type fragments resulting from loss of the reducing-
```

terminal GlcNAc residue. The [M + Na]+ and [M + Li]+ ions from high-mannose and hybrid glycans gave a series of ions of composition (Man)nNa/Li+ where n = 1 to the total number of glycans in the molecule, allowing these sugars to be distinguished from the more highly processed complex glycans. Other ions in the spectra of the high-mannose glycans were diagnostic of chain branching but insufficient information was available to determine their mode of formation.</AbstractText>
  </Abstract>
- <AuthorList>
  <Author>Harvey DJ</Author>
  </AuthorList>
  <PublicationType>Journal Article</PublicationType>
  </PubmedArticle>
- <Structure>
  <ExtendedAscii>a-d-manp-(1-6)+ | a-d-manp-(1-4)-a-d-glcnac-(1-4)-a-d-glcnac | a-d-manp-(1-3)+</ExtendedAscii>
  </Structure>
- <PeakList>
  <MassPeak>138</MassPeak>
- <Peak>
  <MZ>138,0</MZ>
  <Intensity>4,0</Intensity>
  </Peak>
- <Peak>
  <MZ>203,0</MZ>
  <Intensity>8,0</Intensity>
  </Peak>
- <Peak>
  <MZ>226,0</MZ>
  <Intensity>49,0</Intensity>
  </Peak>
- <Peak>
  <MZ>244,0</MZ>
  <Intensity>20,0</Intensity>
  </Peak>
- <Peak>
  <MZ>275,0</MZ>
  <Intensity>8,0</Intensity>
  </Peak>
- <Peak>
  <MZ>329,0</MZ>
  <Intensity>10,0</Intensity>
  </Peak>
- <Peak>
  <MZ>331,1</MZ>
  <Intensity>13,0</Intensity>
  </Peak>
- <Peak>
  <MZ>346,1</MZ>
  <Intensity>25,0</Intensity>
  </Peak>
- <Peak>
  <MZ>347,1</MZ>
  <Intensity>52,0</Intensity>
  </Peak>
- <Peak>
  <MZ>388,1</MZ>
  <Intensity>9,0</Intensity>
  </Peak>
- <Peak>
  <MZ>429,1</MZ>
  <Intensity>17,0</Intensity>

```
  </Peak>
- <Peak>
  <MZ>447,1</MZ>
  <Intensity>32,0</Intensity>
  </Peak>
- <Peak>
  <MZ>493,1</MZ>
  <Intensity>4,0</Intensity>
  </Peak>
- <Peak>
  <MZ>509,1</MZ>
  <Intensity>61,0</Intensity>
  </Peak>
- <Peak>
  <MZ>550,2</MZ>
  <Intensity>54,0</Intensity>
  </Peak>
- <Peak>
  <MZ>583,2</MZ>
  <Intensity>15,0</Intensity>
  </Peak>
- <Peak>
  <MZ>629,2</MZ>
  <Intensity>9,0</Intensity>
  </Peak>
- <Peak>
  <MZ>712,2</MZ>
  <Intensity>100,0</Intensity>
  </Peak>
- <Peak>
  <MZ>730,2</MZ>
  <Intensity>12,0</Intensity>
  </Peak>
- <Peak>
  <MZ>771,2</MZ>
  <Intensity>8,0</Intensity>
  </Peak>
- <Peak>
  <MZ>832,2</MZ>
  <Intensity>5,5</Intensity>
  </Peak>
- <Peak>
  <MZ>915,3</MZ>
  <Intensity>7,0</Intensity>
  </Peak>
- <Peak>
  <MZ>933,3</MZ>
  <Intensity>20,0</Intensity>
  </Peak>
  </PeakList>
  </SweetDBMassSpectra>
```

11.2 XML-Format zum Austausch von NMR-Spektren

```
  <?xml version="1.0" encoding="utf-8" ?>
  <!DOCTYPE SweetDBNMRSpectra (View Source for full doctype...)>
- <SweetDBNMRSpectra>
- <User>
  <UserID>0</UserID>
  <UserName>kerschgens</UserName>
  </User>
```

```
- <PubmedArticle>
  <PMID>12234251</PMID>
  <Year>2003</Year>
- <Journal>
  <JournalName>Biochem J</JournalName>
  <Volume>369</Volume>
  <Issue />
- <PubDate>
  <Year>2003</Year>
  </PubDate>
  </Journal>
  <ArticleTitle>Structural elucidation of zwitterionic carbohydrates
derived from glycosphingolipids of the porcine parasitic nematode
Ascaris</ArticleTitle>
- <Pagination>
  <MedlinePgn>89-102</MedlinePgn>
  </Pagination>
- <Abstract>
  <AbstractText>Carbohydrates substituted with phosphocholine (PC) and
phosphoethanolamine (PE) were released from zwitterionic glycosphingolipids
of the pig parasitic nematode Ascaris suum by treatment with
endoglycoceramidase. Individual glycans were obtained by HPLC on porous
graphitic carbon followed by high-pH anion-exchange chromatography. In
addition to the known pentasaccharides Gal alpha 3GalNAc beta 4[PC6]GlcNAc
beta 3Man beta 4Glc and Gal alpha 3GalNAc beta 4[PC6]GlcNAc beta 3[PE6]Man
beta 4Glc, the corresponding tri- and tetra-saccharides, as well as
components with elongated structures, could be identified by matrix-
assisted laser-desorption ionization-time-of-flight MS, methylation
analysis, 1H- and 13C-NMR spectroscopy, exoglycosidase cleavage and
electrospray ionization ion-trap MS. The extended components comprised
novel structural motifs such as di-substituted alpha-galactose carrying two
beta-linked galactosyl residues, which were found to bear, in part, further
fucose, galactose, N -acetylgalactosamine and/or N -acetylglucosamine
moieties. Furthermore, additional fucosylation of the PC-substituted N -
acetylglucosamine and a non-terminal fucosyl motif were detected. In
conclusion, this study contributes significant new information on the
glycome of nematodes.</AbstractText>
  </Abstract>
- <AuthorList>
  <Author>Friedl CH, Lochnit G, Zahringer U, Bahr U, Geyer R</Author>
  </AuthorList>
  <PublicationType>Journal Article</PublicationType>
  </PubmedArticle>
- <Structure>
  <ExtendedAscii>PC-(3-6)+ | a-D-Galp-(1-3)-b-D-GalpNAc-(1-4)-b-D-GlcpNAc-
(1-3)-b-D-Manp-(1-4)-a-D-Glcp</ExtendedAscii>
  </Structure>
- <Specimen>
  <Name>Zwitterionic carbohydrates of the porcine parasitc nematode Ascaris
suum [Component A; a-D-Glcp]</Name>
  <Class>Carbohydrate</Class>
  <Spectra>Proton NMR</Spectra>
  </Specimen>
- <Measurement>
- <Equipment>
  <Spectrometer>Avance DRX 600 Bruker spectrometer</Spectrometer>
  <Operator />
  <MHz />
  </Equipment>
- <Conditions>
  <Temperature>300</Temperature>
  <Solvent>©˜H2O</Solvent>
```

```
  </Conditions>
  </Measurement>
- <BiologicalSource>
  <Species>Ascaris suum</Species>
  <ProteinName />
  <System />
  <Disease />
  <SwissProtID />
  <Comment />
  </BiologicalSource>
- <PeakList>
- <Peak>
  <Nuclei>C</Nuclei>
  <Atom>C</Atom>
  <Residue>A-D-GLCP</Residue>
  <Linkage />
  <PPM>1111</PPM>
  </Peak>
.
.
.

- <Peak>
  <Nuclei>H</Nuclei>
  <Atom>H12</Atom>
  <Residue>PC</Residue>
  <Linkage>4, 3, 6</Linkage>
  <PPM>0</PPM>
  </Peak>
  </PeakList>
  </SweetDBNMRSpectra>
```

CarbStore

Manual
V 1.0 (Mai 2003)

by

Klaus Lohmann (k.lohmann@dkfz.de)
Claus-W. von der Lieth (w.vonderlieth@dkfz.de)
German Cancer Research Center
Im Neuenheimer Feld 280
69120 Heidelberg
Tel: 0049-6221-424541
Fax:0049-6221-424554
Germany

Philosophy
The basic philosophy of **CarbStore** is to encourage scientists in the area of glycobiology to make their original scientific data like MS- and NMR-spectra accessible to the public using this service. Often original scientific data are not documented in printed publications and can not be accessed by the scientific community. **CarbStore** is web-based interface designed to input carbohydrate related original scientific data and information. The data will first be stored in private areas – each scientist will maintain his/her own database – and should be released to the publicly available database ***SWEET-DB*** at a suitable time. It is up to each individual scientist to decide when and which entries shall be released to ***SWEET-DB***.
Purpose

The main purpose of **CarbStore** is to provide an efficient way to input, annotate and store experimental original data like MS- and NMR-spectra as well as reference data. **CarbStore** can be used to store private data, which can be easily linked with the existing *SWEET-DB*. It is the decision of each user when to make each entry available to the scientific data public. The current description does not include the input of NMR-spectra.. An additional description is available for the input of NMR spectra.

This is a beta version of **CarbStore**. We are asking for your help to improve the service so that we can develop it along the needs of the scientific community. In case you find any problems using **CarbStore** please do not hesitate to contact us. Also we would appreciate any comment or hint so that we can make the service working more convenient according your needs.

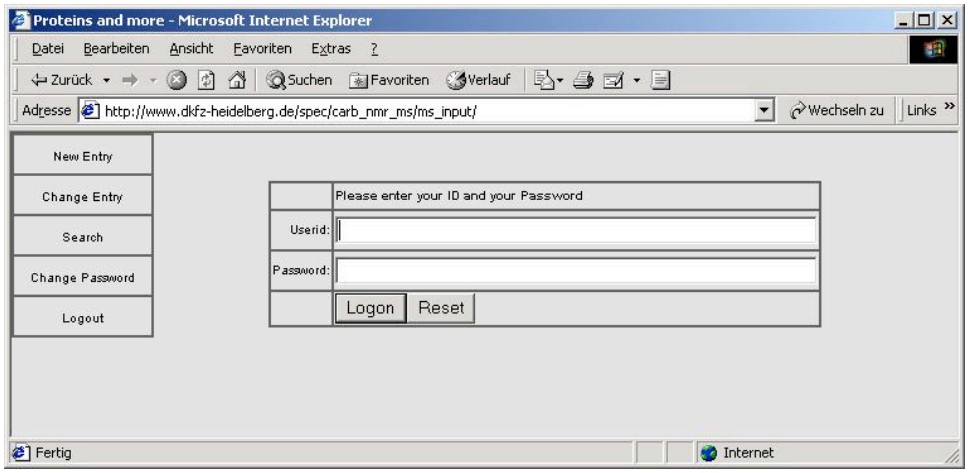
Please contact Klaus Lohmann (k.lohmann@dkfz-heidelberg.de) or
Dr. Claus-W. von der Lieth (w.vonderlieth@dkfz-heidelberg.de),
German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.

WWW
You can reach **CarbStore** by entering
http://www.dkfz-heidelberg.de/spec/carb_nmr_ms/ms_input/
in the address field of your browser.

This service has been tested with Internet Explorer 5.5 and 6. as well as Netscape 4.7.. It is recommended to use one of these browsers.

Login

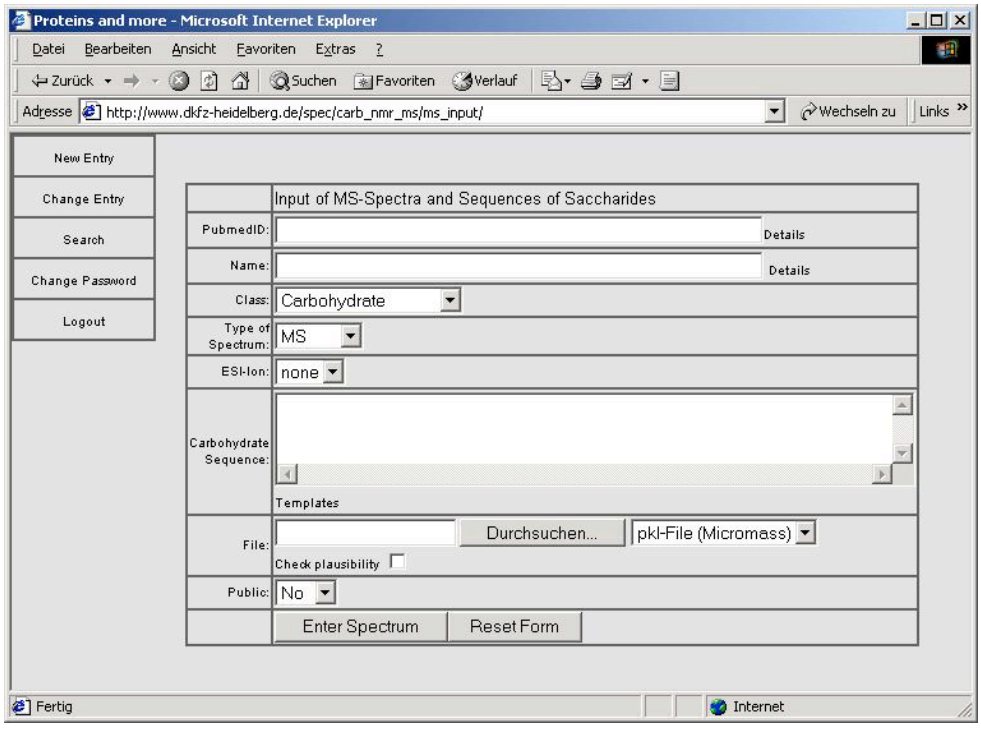
Each user will have a private database which can only be accessed by the authorized persons. Therefore, entering new data or changing of already stored data is only possible if you identify yourself with a user-id and a password. Both can be obtained by sending an email to k.lohmann@dkfz-heidelberg.de. After you have received your user id and password, you have to input both in the following form:



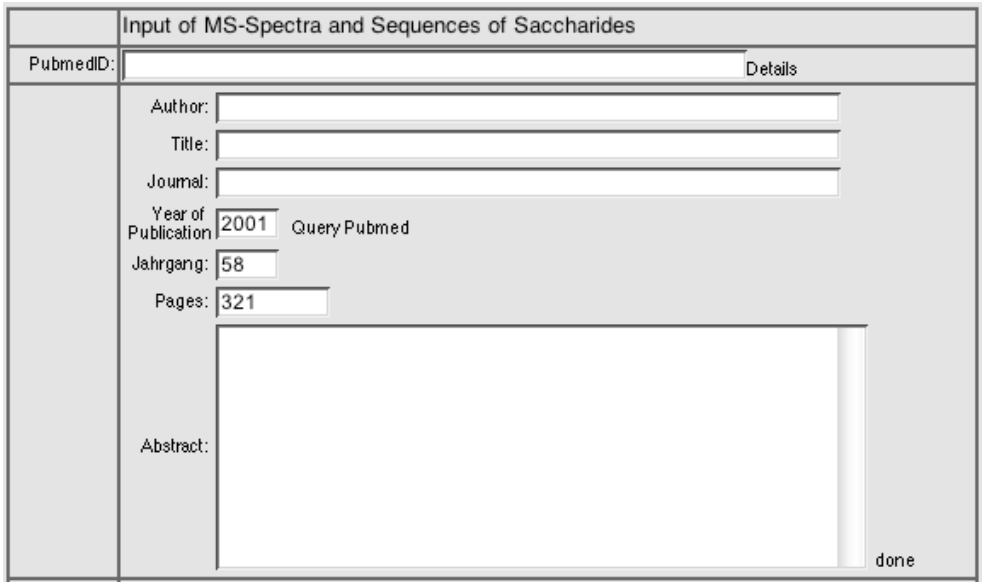
If user-id and password are correctly entered you can use all options to input, change and search data as indicated in the menu areas on the left side of the screen. You can select one following items:

a) *New Entry*: enter a new set of data
b) *Change Entry* : list and change already stored data
c) *Search*: search ***SWEET-DB*** and/or your own database
d) *Change Password*, this item gives you the opportunity to change the password, so that no one else can have a look at your data
e) *Logout*: end your session

Enter a new entry with spectrum
Activating ‘New Entry’, you receive the following input form:

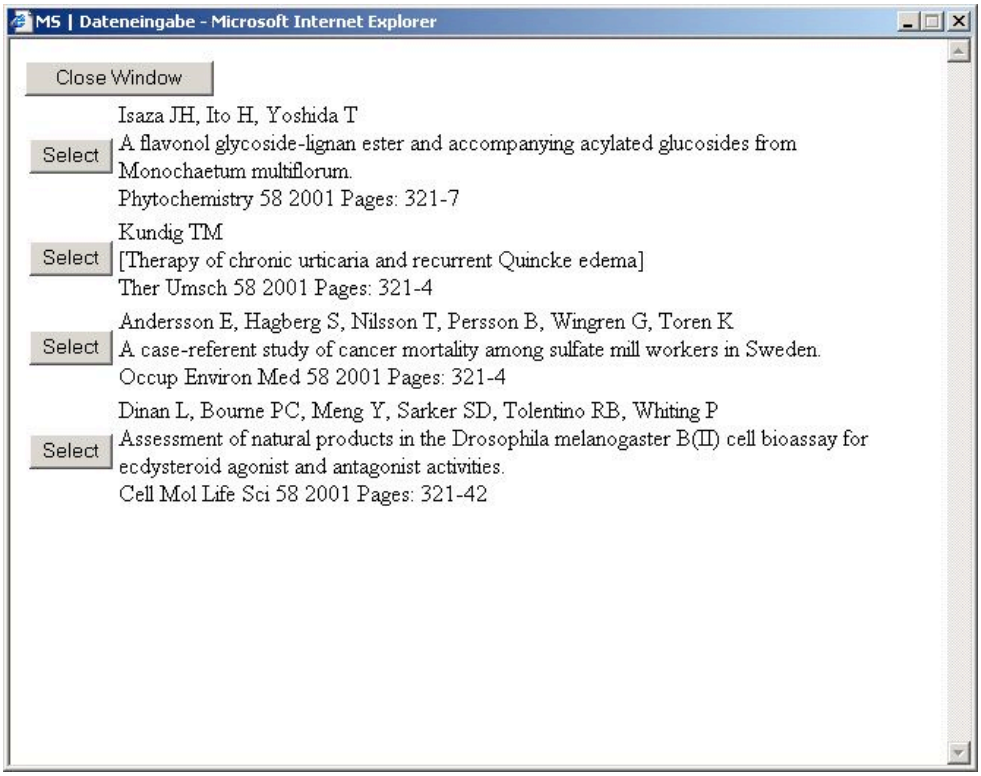


Entering Reference Data (Pubmed-ID)
To make the input of references as simple as possible, we establish an direct link to the NCBI-PubMed-Service. Only the PubMedID is required to input bibliographic data including complete abstracts. To retrieve the PubMedID click on the item *details* in the grey area in the PubMedID line.The following input spreadsheet will appear:

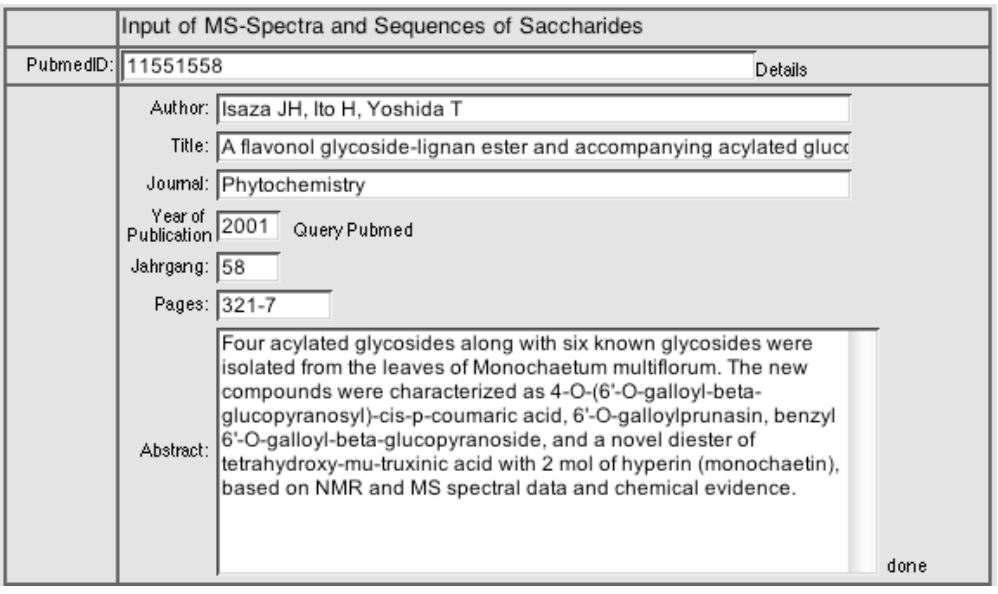


The best way to search for a **Pubmed**-ID is to look for numerical data like *Year of Publication*, *Pages*, Volume and one or two authors. Input your specific data in these field and send the query to PubMed (click on the Text Ouery Pubmed). Normally,

only a few entries will be retrieved. A new window pops up and presents a list of all papers fitting the query.



The next step is to select the desired entry by pressing the correspoinding ***SELECT***-button. All data given with the PubMed-entry will be automatically entered into the input spreadsheet (see below) . In case you could not find the desired entry, you close the window by hitting the ***CLOSE WINDOW***-button and you have to add all data manually into the reference input form. To store the bibliographical data click the ***done*** link (bottom right).



Entering the Compound-Data

To input the data that identifies your compound, hit the *details*-link in the field Name. There are three fields to be filled with data.

a) Name: Enter the name of the compound
b) Origin: Enter the origin from the compound (e.g.: the species)
c) Tissue: Enter the tissue where the compound has been identified (e.g.: leaves, roots, brain, …)

To finish input of data push the *done*-button to go back to the main input spread sheet.

Class
Enter here the class of your compound. You can choose between *carbohydrate* that means your structure is a 'pure' carbohydrate, or glycosylated protein, which means your compound consist of a glycan attached to a protein (or a part of it). If you choose glycosylated protein the input spreadsheet form will provide two additional input fields.

Protein Sequence
When you choose *glycosylated protein* you enter the protein sequence using the standard one letter for amino acids. **In the field** Glycosylated residue you enter the number of residues to which carbohydrates are covalently linked.

ESI-Ion
Here you can enter the esi-ion, which was used to record the MS-spectra.
Entering the Carbohydrate Sequence

To input your carbohydrate structure the so-called extended form to describe Oligosaccharides is used. The definition can be found at

http://www.chem.qmw.ac.uk/iupac/2carb/38.html

Example of a N-Glycan structure

```
          a-D-Manp-(1-2)-a-D-Manp-(1-6)+
                                       |
                                a-D-Manp-(1-6)+
                                       |      |
          a-D-Manp-(1-2)-a-D-Manp-(1-3)+      |
                                              |
                                   b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc
                                              |
a-D-Manp-(1-2)-a-D-Manp-(1-2)-a-D-Manp-(1-3)+
```

Using the input window a few special input rules have to be fulfilled which are discussed for the N-Glycan structure given above.

Additional Input rules

a) a,b have to used instead of □,□.
b) the ring configuration (D or L) has to be given always.
c) a hyphen before and after the linkage information is required –(1-3)-
d) Furanoses and Pyranoses have to be indicated.

e) A '+'-sign is required to indicate continuation of the structure in an upper of lower line of the editor.
f) To indicate the vertical connections the '+','|', and the connection to the corresponding monosaccharide have to be in one vertical line. This requirement often causes problems when inputting structures by the copy/paste option when a proportional fond (like Arial or Times) is used.

Correct input

```
a-D-Manp-(1-2)-a-D-Manp-(1-6)+
                              |
                          a-D-Manp-(1-6)
                              |
 a-D-Manp-(1-2)-a-D-Manp-(1-3)+
```

Incorrect input:

```
a-D-Manp-(1-2)-a-D-Manp-(1-6)+
                              |
                          a-D-Manp-(1-6)
                              |
  a-D-Manp-(1-2)-a-D-Manp-(1-3)+
```

Incorrect input:

```
a-D-Manp-(1-2)-a-D-Manp-(1-6)+
                              |
                          a-D-Manp-(1-6)
                              |
            a-D-Manp-(1-2)-a-D-Manp-(1-3)+
```

A list of common carbohydrate structures can be found at

http://www.dkfz-heidelberg.de/spec/sweet2/doc/sam/leftframe.html

To test, if the input of a residue is conform to the nomenclature activate

http://www.dkfz-heidelberg.de/spec/sweet2/doc/input/testtemplate.html

Uploading the file containing the spectrum.
Using the field 'file' you can upload the file containing the spectrum: hit the button *SEARCH* (Depends of your language and the browser you use, normally it is the button beside the textfield) and you will be prompted for a file containing the MS-spectra. It will be uploaded and stored in the database. **CarbStore** is currently able to store files in Micromass pkl-format and Unidat cdf-format, which is internally converted to pkl-format. **CarbStore** will check your data for plausibility. Using the ***GLYCO-FRAGMENT*** tool (www.dkfz.de/spec/glycofragment/) a theoretical fragmentation scheme will be calculated based from the carbohydrate structure you have input. The calculated peaks will be compare with the data contained in the uploaded file. At least 60 percent of the theoretically expected peaks should be there. If this is not the case, you will get a warning message. Please check, that the

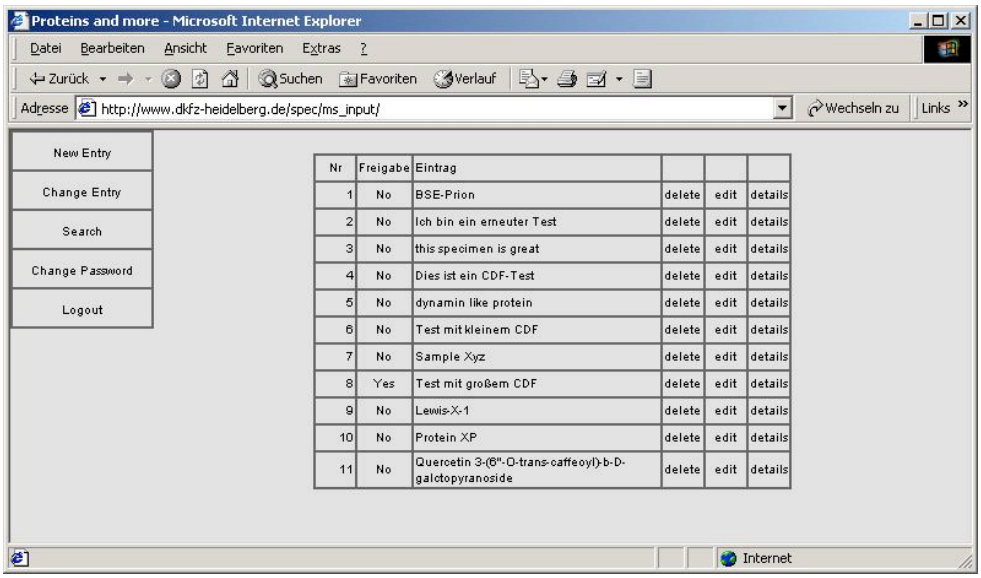
uploaded file an the structure correspond to each other. In case you are sure, that structure and spectrum are both correct, you may ignore the warning message and the MS spectrum data will be stored assigned with the corresponding carbohydrate structure..

Making your data available to the public

In case you want to make your data available to the public, simply choose ‘yes’ from the pull down menu in the field ‘Public’. All the data you input will be transferred to the ***SWEET-DB***. In case you want to do this later, active ‘no’ in the public field. To make your data publicly available at a later time, activate the ‘Change Entry ’ option from the main menu an change the ‘no’ in the public field to ‘yes’. Thus the user can control the status of each entry at any time. All data of one entry keep private as long as the user has not agreed to make it available to the public.

Annotate your data

Of course it is possible to annotate data that are already stored in **CarbStore**. Simply click '*CHANGE ENTRY*' from the menu list on the left side:



Nr	Freigabe	Eintrag			
1	No	BSE-Prion	delete	edit	details
2	No	Ich bin ein erneuter Test	delete	edit	details
3	No	this specimen is great	delete	edit	details
4	No	Dies ist ein CDF-Test	delete	edit	details
5	No	dynamin like protein	delete	edit	details
6	No	Test mit kleinem CDF	delete	edit	details
7	No	Sample Xyz	delete	edit	details
8	Yes	Test mit großem CDF	delete	edit	details
9	No	Lewis-X-1	delete	edit	details
10	No	Protein XP	delete	edit	details
11	No	Quercetin 3-(6''-O-trans-caffeoyl)-b-D-galctopyranoside	delete	edit	details

All entries you have input so far will be displayed.

Deleting a spectrum

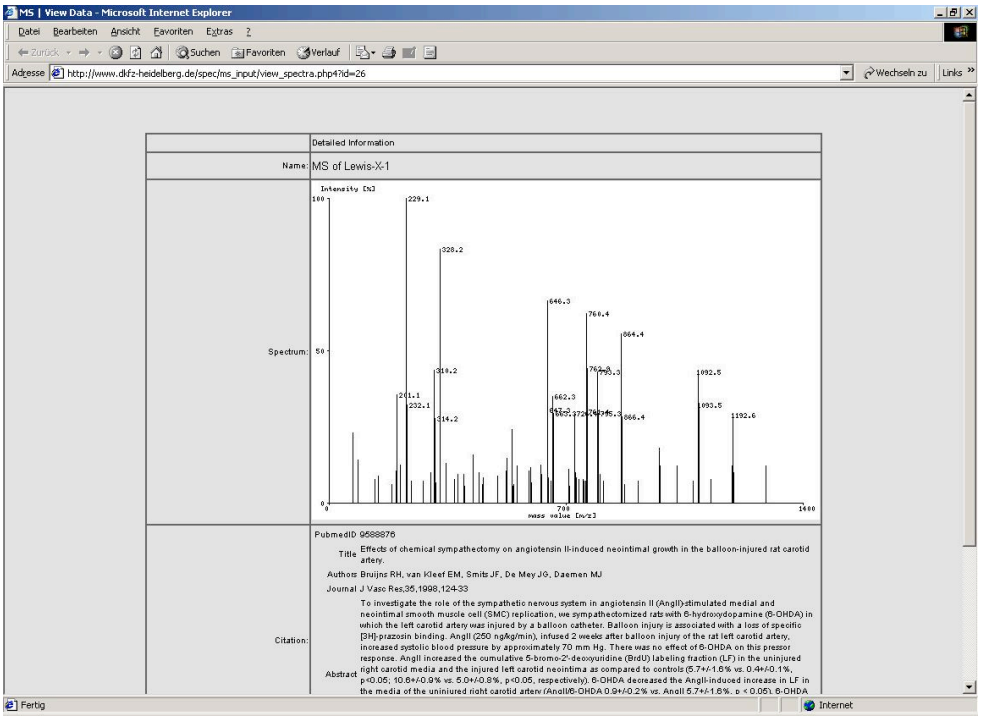
If you want to delete a stored spectrum just hit the ***DELETE***-link. After confirming the action the spectrum is permanently removed from the database.

Editing a spectrum

By hitting the ***EDIT***-link you are able to change any of the stored data field. The input spreadsheet is identical with the one for entering new entries. When you have finished just hit the ***SAVE CHANGES*** –button and the changes will be updated in the databases.

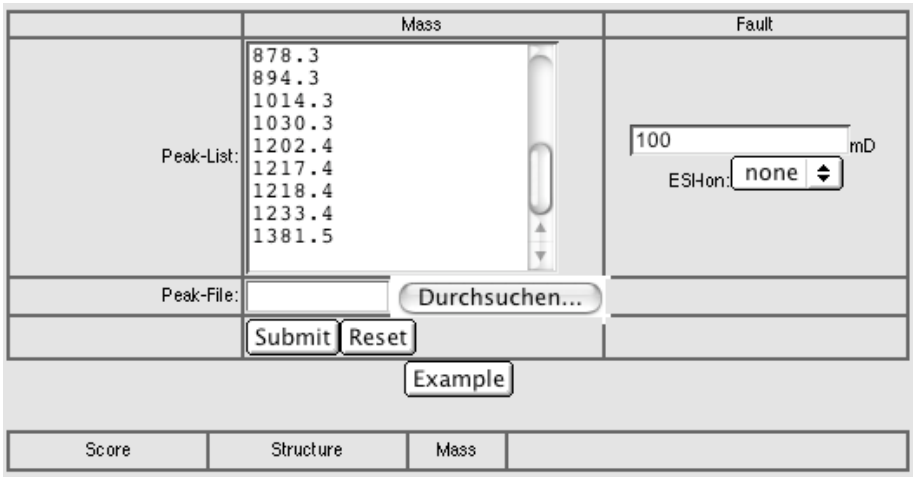
Viewing details

To get a detailed view of the stored data you hit the link ***DETAILS.*** In a new window you see a printable view of the data, including a plotted spectra:

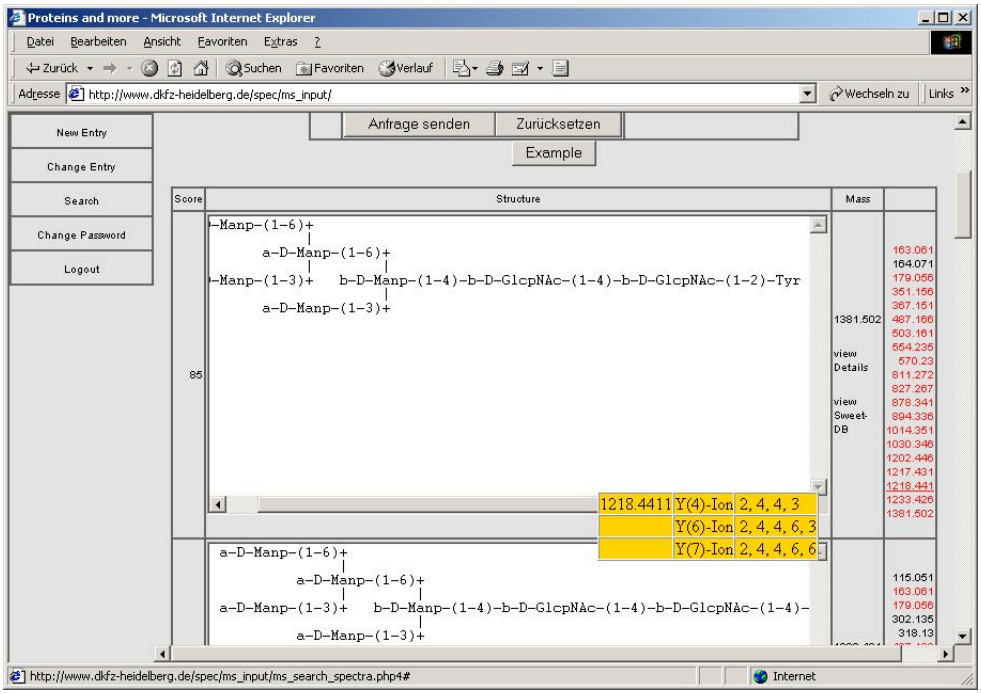

MS | View Data - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Suchen Favoriten Verlauf
Adresse http://www.dkfz-heidelberg.de/spec/ms_input/view_spectra.php4?id=26
Wechseln zu Links
Detailed Information
Name: MS of Lewis-X-1
Spectrum:
Intensity [%]
mass value [m/z]
Citation:
PubmedID 9588876
Title Effects of chemical sympathectomy on angiotensin II-induced neointimal growth in the balloon-injured rat carotid artery.
Authors Bruijns RH, van Kleef EM, Smits JF, De Mey JG, Daemen MJ
Journal J Vasc Res.35,1998,124-33
Abstract
To investigate the role of the sympathetic nervous system in angiotensin II (AngII)-stimulated medial and neointimal smooth muscle cell (SMC) replication, we sympathectomized rats with 6-hydroxydopamine (6-OHDA) in which the left carotid artery was injured by a balloon catheter. Balloon injury is associated with a loss of specific [3H]-prazosin binding. AngII (250 ng/kg/min), infused 2 weeks after balloon injury of the rat left carotid artery, increased systolic blood pressure by approximately 70 mm Hg. There was no effect of 6-OHDA on this pressor response. AngII increased the cumulative 5-bromo-2'-deoxyuridine (BrdU) labeling fraction (LF) in the uninjured right carotid media and the injured left carotid neointima as compared to controls (5.7+/-1.6% vs. 0.4+/-0.1%, p<0.05; 10.6+/-0.9% vs. 5.0+/-0.8%, p<0.05, respectively). 6-OHDA decreased the AngII-induced increase in LF in the media of the uninjured right carotid artery (AngII/6-OHDA 0.9+/-0.2% vs. AngII 5.7+/-1.6%, p < 0.05). 6-OHDA
Fertig
Internet

Search for a Spectra

If you want to search for spectra that are already stored in ***SWEET-DB*** you can hit the ***SEARCH*** link in the menu and you will see the following form:



Enter a list of peaks manually or upload a spectrum that is stored on your local hard disk. An example is provided to demonstrate the procedure. Please hit the ***EXAMPLE***-button. A list of peaks will be entered automatically into the text field. Hit the ***SUBMIT*** Button The **SWEET_DB** MS-database of theoretically calculated MS-spectra will be searched. A list of best matching entries will be displayed in descending order of the number of matches. A score of 100 means that all input peaks have been matched.

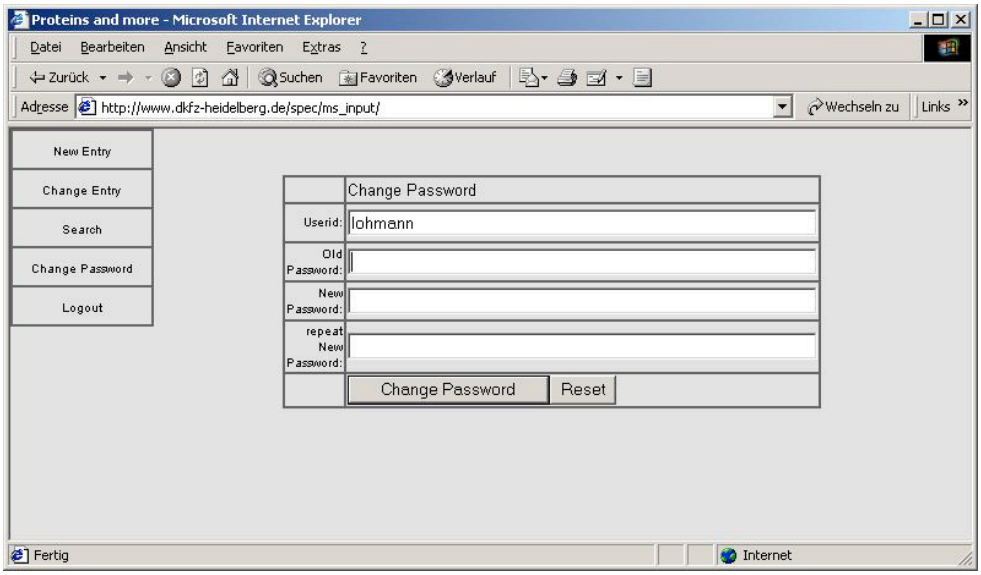


Each result line consists of the score, the structure, the mass, and the peak list of the matched spectrum. To provide a more detailed look at the fragmentation scheme of the structure one may hit the VIEW DETAILS button. A link the Glyco-MS_FRAGMENT will be established and will allow to analyse the fragmentation scheme in detail. In case one wants to see the additional data stored for this entry in

SWEET-DB, hit the VIEW SWEETDB button. A direct link the *SWEET-DB* for this entry will be established. Moving the mouse over the peak list at the right side the fragmentation will be displayed which gives rise to that specific masst.

Changing your password

To protect your data in the best way you can change your password. You should do this directly after you receive your password from us, and from time to time. You can do this by clicking on the link *CHANGE PASSWORD*. You see this form:

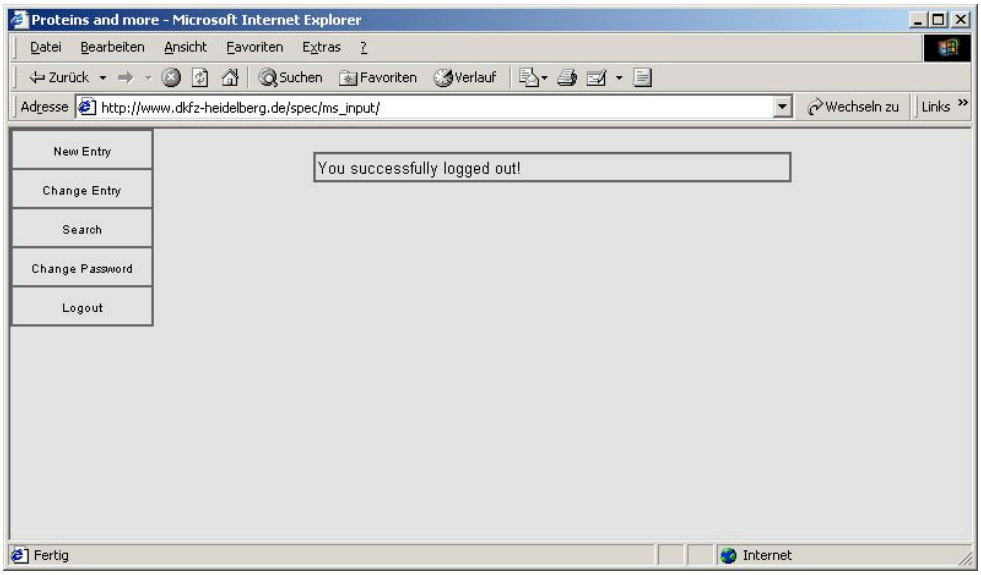


Just enter your old password and for security reason two times the new one. Hit the *CHANGE PASSWORD* button, and you have successfully changed your password.

Finishing your Session

After you have finished entering or changing your data please hit the *LOGOUT* link to prevent any other from manipulating your data. You are logged out after sixty minutes for security reasons.

If you have successfully logged out you see this mesagge:



12 Appendix A

12.1 Extended Description for Oligosaccharides

The numbers following the monosaccharide description indicate the location of the substitution at the corresponding C-Atom of the sugar.

a-D-Galp2OMe, a-D-Galp2OAcOAc, a-Galp2OMe3OMe,4OMe,6OMe

12.2 List of currently supported substituents

Input code	Substituent
N	amine
Fluoro	fluorine
Chloro	chlorine
Bromo	bromine
Me	methyl
Ac	acetate
NAC	N-Acetyl
OAc	O-Acetyl
OMe	O-Methyl
SO3	sulfite
OSO3	sulfate
NSO3	N-sulfate
PO4	phosphate
Gc	glycolic acid
CH	Choline
EA	Ethanolamine
PA	2-Aminopyridine
Py	Pyridine
PCH	Phosphatidyl Choline
PEA	Phosphatidyl Ethanolamine

Tryptische Fragmente von **Dynamin like protein.- Dictyostelium discoideum (Slime mold)**
Masse 96047

MDQLIPVINK
LQDVFNTLGSDPLDLPQIVVVGSQSSGK
SSVLENIVGR
DFLPR
GSGIVTR
R
PLILQLTHLPIADDGSQTQEWGEFLHK
PNDMFYDFSEIR
EEIIR
DTDR
MTGK
NK
GISAQPINLK
IYSPHVVNLTLVDLPGITK
VPVGDQPTDIEQQIR
R
MVMAYIK
K
QNAIIVAVTPANTDLANSDALQLAK
EVDPEGK
R
TIGVITK
LDLMDK
GTDAMEVLTGR
VIPLTLGFIGVINR
SQEDIIAK
K
SIR
ESLK
SEILYFK
NHPIYK
SIANR
SGTAYLSK
TLNK
LLMFHIR
DTLPDLK
VK
VSK
MLSDVQGELSTYGDPLYDTK
NSQGALLLQIITIFSSNFK
DAIDGK
LTDLSNNELYGGAR
ISYIFNEIYSHCVNNIDPLEGISLNDIR
TTMR
NATGPR

AALFIPEISFELLVK
K
QVVR
LEEPSAQCVEYVYDELQR
IVSQLEAK
ELSR
FINLK
AR
VIEVVNNLLQK
HK
VPTK
TMIEHLIK
IETAFINTSHPDFVGGEGIFESLYK
K
QQLQQQNHLQQLQDQYQQQQQQQQQQQQQNGINNNQK
GDNGNMNVNQQNMNQQNMNQQNQSTNPFLQQQQQGQNK
YPGGPPAQQQPNQQPNQLNK
GPQNMPPNQSK
PSSIPQNGPNNNNNNNNNNR
QDHQQGSFFSSFFR
ASPDPSLGQYGGANNSNNSNNPTSPINSSSNSGNNYNTFGGQQSSSSSSQQLQQSSQS
QYK
TSYNNNNNSSSNNSSYNR
YQDDFYGR
GDK
LNQVPSIIK
APDDLTSK
EK
FETELIR
ELLISYFNIVK
K
NVK
DSVPK
SIMHFLVNQSK
EHIQNELVAALYK
EELFDELLEESPQISSK
R
K
SCK
AMIEILR
K
ANEIINEIR
DFR
N

13 Literaturverzeichnis

1. Werning, C., *Medizin für Apotheker*. 1 ed. 1987: Wissenschaftliche Verlagsgesellschaft. 790.
2. Wichtl, M., *Teedrogen*. 2 ed. 1989: Wissenschaftliche Verlagsgesellschaft. 568.
3. Hänsel, R., *Pharmakognosie, Phytopharmazie*. 7 ed. 2003. 1403.
4. Taylor, M.E. and K. Drickamer, *Introduction to Glycobiology*. 2002: Oxford University Press. 224.
5. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
6. Lottspeich, F. and H. Zorbas, *Bioanalytik*. 1998: Spektrum Akademischer Verlag, Heidelberg - Berlin. 1035.
7. Ritchie, G.E., et al., *Glycosylation and the complement system.* Chem Rev, 2002. **102**(2): p. 305-20-19.
8. Sinnott, M.L., *Catalytic mechanisms of enzymatic glycosyl transfer.* Chem Rev, 1990. **90**: p. 1171-1202.
9. Coutinho, P. and B. Henrissat, *Carbohydrate-active enzymes: an integrated database approach.* Recent Advances in Carbohydrate Bioengineering, 1999: p. 3-12.
10. Coutinho, P. and B. Henrissat, *The modular structure of cellulases and other carbohydrate-active enzymes: an integrated database approach.* Genetics, Biochemistry and Ecology of Cellulose Degradation, 1999: p. 15-23.
11. Linn, T.C., F.H. Pettit, and L.J. Reed, *Alpha-keto acid dehydrogenase complexes. X. Regulation of the activity of the pyruvate dehydrogenase complex from beef kidney mitochondria by phosphorylation and dephosphorylation.* Proc Natl Acad Sci U S A, 1969. **62**(1): p. 234-41.
12. Ingebritsen, T.S., *Protein phosphorylation and the hormonal control of hepatic cholesterol synthesis.* Biochem Soc Trans, 1983. **11**(6): p. 644-6.
13. Alessi, D.R., et al., *PD 098059 is a specific inhibitor of the activation of mitogen-activated protein kinase kinase in vitro and in vivo.* J Biol Chem, 1995. **270**(46): p. 27489-94.
14. Favata, M.F., et al., *Identification of a novel inhibitor of mitogen-activated protein kinase kinase.* J Biol Chem, 1998. **273**(29): p. 18623-32.
15. Sebolt-Leopold, J.S., et al., *Blockade of the MAP kinase pathway suppresses growth of colon tumors in vivo.* Nat Med, 1999. **5**(7): p. 810-6.
16. Lee, J.C., et al., *Bicyclic imidazoles as a novel class of cytokine biosynthesis inhibitors.* Ann N Y Acad Sci, 1993. **696**: p. 149-70.
17. Kadmon, G. and P. Altevogt, *The cell adhesion molecule L1: species- and cell-type-dependent multiple binding mechanisms.* Differentiation, 1997. **61**(3): p. 143-50.
18. Clark, R.A., et al., *Characterisation of tissue-specific oligosaccharides from rat brain and kidney membrane preparations enriched in Na+,K+-ATPase.* Glycoconj J, 1999. **16**(8): p. 437-56.
19. Benlagha, K., et al., *Modifications of Igalpha and Igbeta expression as a function of B lineage differentiation.* J Biol Chem, 1999. **274**(27): p. 19389-96.

20. Rudd, P.M., et al., *Glycosylation differences between the normal and pathogenic prion protein isoforms.* Proc Natl Acad Sci U S A, 1999. **96**(23): p. 13044-9.
21. Bohne, A. and C.W. von der Lieth, *Glycosylation of proteins: a computer based method for the rapid exploration of conformational space of N-glycans.* Pac Symp Biocomput, 2002: p. 285-96.
22. Chester, M.A., *IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature of glycolipids--recommendations 1997.* Eur J Biochem, 1998. **257**(2): p. 293-8.
23. McNaught, A.D., *Nomenclature of carbohydrates (recommendations 1996).* Adv Carbohydr Chem Biochem, 1997. **52**: p. 43-177.
24. Zhang, Y., et al., *A novel monoantennary complex-type sugar chain found in octopus rhodopsin: occurrence of the Gal beta1-->4Fuc group linked to the proximal N-acetylglucosamine residue of the trimannosyl core.* Glycobiology, 1997. **7**(8): p. 1153-8.
25. Taguchi, T., et al., *Proton NMR study of the trimannosyl unit in a pentaantennary N-linked decasaccharide structure. Complete assignment of the proton resonances and conformational characterization.* Eur J Biochem, 1995. **228**(3): p. 822-9.
26. Taguchi, T., et al., *Complete assignments of 13C NMR resonances to all the carbon atoms of the trimannosido-di-N-acetylchitobiosyl structure in a pentaantennary decasaccharide glycopeptide.* Carbohydr Res, 1995. **275**(1): p. 185-91.
27. Zachara, N.E. and G.W. Hart, *The emerging significance of O-GlcNAc in cellular regulation.* Chem Rev, 2002. **102**(2): p. 431-8.
28. Mutschler, E., *Arzneimittelwirkungen*. 8 ed. 2001: Wissenschaftliche Verlagsgesellschaft. 1186.
29. Fromtling, R. and J. Castañer, *Zanamivir. Antiviral, Influenza neuraminidase inhibitor.* Drugs of the future, 1996. **21**(4): p. 375-382.
30. Bardsley-Elliot, A. and S. Noble, *Oseltamivir.* Drugs, 1999. **58**(5): p. 851-60; discussion 861-2.
31. von Itzstein, M., et al., *A study of the active site of influenza virus sialidase: an approach to the rational design of novel anti-influenza drugs.* J Med Chem, 1996. **39**(2): p. 388-91.
32. von Itzstein, M., et al., *Rational design of potent sialidase-based inhibitors of influenza virus replication.* Nature, 1993. **363**(6428): p. 418-23.
33. Kiefel, M.J. and M. von Itzstein, *Recent advances in the synthesis of sialic acid derivatives and sialylmimetics as biological probes.* Chem Rev, 2002. **102**(2): p. 471-90.
34. Mattu, T.S., et al., *O-glycan analysis of natural human neutrophil gelatinase B using a combination of normal phase-HPLC and online tandem mass spectrometry: implications for the domain organization of the enzyme.* Biochemistry, 2000. **39**(51): p. 15695-704.
35. Guile, G.R., et al., *A rapid high-resolution high-performance liquid chromatographic method for separating glycan mixtures and analyzing oligosaccharide profiles.* Anal Biochem, 1996. **240**(2): p. 210-26.
36. Sagi, D., et al., *Sequencing of tri- and tetraantennary N-glycans containing sialic acid by negative mode ESI QTOF tandem MS.* J Am Soc Mass Spectrom, 2002. **13**(9): p. 1138-48.

37. Satomaa, T., et al., *O-glycans on human high endothelial CD34 putatively participating in L-selectin recognition.* Blood, 2002. **99**(7): p. 2609-11.
38. Mechref, Y. and M.V. Novotny, *Structural investigations of glycoconjugates at high sensitivity.* Chem Rev, 2002. **102**(2): p. 321-69.
39. Ekeberg, D., S.H. Knutsen, and M. Sletmoen, *Negative-ion electrospray ionisation-mass spectrometry (ESI-MS) as a tool for analysing structural heterogeneity in kappa-carrageenan oligosaccharides.* Carbohydr Res, 2001. **334**(1): p. 49-59.
40. Monsarrat, B., et al., *Characterization of mannooligosaccharide caps in mycobacterial lipoarabinomannan by capillary electrophoresis/electrospray mass spectrometry.* Glycobiology, 1999. **9**(4): p. 335-42.
41. Dancik, V., et al., *De novo peptide sequencing via tandem mass spectrometry.* J Comput Biol, 1999. **6**(3-4): p. 327-42.
42. Mo, W., et al., *Structural analysis of oligosaccharides derivatized with 4-aminobenzoic acid 2-(diethylamino)ethyl ester by matrix-assisted laser desorption/ionization mass spectrometry.* Anal Chem, 1998. **70**(21): p. 4520-6.
43. Nemeth, J.F., et al., *Characterization of the glycosylation sites in cyclooxygenase-2 using mass spectrometry.* Biochemistry, 2001. **40**(10): p. 3109-16.
44. Schlosser, A., et al., *Analysis of post-translational modification and characterization of the domain structure of dynamin A from Dictyostelium discoideum.* J Mass Spectrom, 2003. **38**(3): p. 277-82.
45. Schlosser, A., et al., *Identification of protein phosphorylation sites by combination of elastase digestion, immobilized metal affinity chromatography, and quadrupole-time of flight tandem mass spectrometry.* Proteomics, 2002. **2**(7): p. 911-8.
46. Macek, B., J. Hofsteenge, and J. Peter-Katalinic, *Direct determination of glycosylation sites in O-fucosylated glycopeptides using nano-electrospray quadrupole time-of-flight mass spectrometry.* Rapid Commun Mass Spectrom, 2001. **15**(10): p. 771-7.
47. Hui, J.P., T.C. White, and P. Thibault, *Identification of glycan structure and glycosylation sites in cellobiohydrolase II and endoglucanases I and II from Trichoderma reesei.* Glycobiology, 2002. **12**(12): p. 837-49.
48. Schlack, P. and W. Kumpf, *Über eine Methode zur Ermittlung der Konstitution von Peptiden.* Z. Physiol. Chemie Hoppe-Seyler, 1926. **154**: p. 125-170.
49. Inglis, A.S., *Chemical procedures for C-terminal sequencing of peptides and proteins.* Anal Biochem, 1991. **195**(2): p. 183-96.
50. Edman, P. and G. Begg, *A protein sequenator.* Eur J Biochem, 1967. **1**(1): p. 80-91.
51. Peracaula, R., et al., *Glycosylation of human pancreatic ribonuclease: differences between normal and tumour states.* Glycobiology, 2002.
52. Suzuki, S. and S. Honda, *A tabulated review of capillary electrophoresis of carbohydrates.* Electrophoresis, 1998. **19**(15): p. 2539-60.
53. Bigge, J.C., et al., *Nonselective and efficient fluorescent labeling of glycans using 2-amino benzamide and anthranilic acid.* Anal Biochem, 1995. **230**(2): p. 229-38.
54. Harvey, D.J., *Identification of protein-bound carbohydrates by mass spectrometry.* Proteomics, 2001. **1**(2): p. 311-28.

55. Garozzo, D., et al., *Discrimination of isomeric oligosaccharides and sequencing of unknowns by post source decay matrix-assisted laser desorption/ionization time-of-flight mass spectrometry.* Rapid Commun Mass Spectrom, 1997. **11**(14): p. 1561-6.
56. Gaucher, S.P., et al., *Mass spectral characterization of lipooligosaccharides from Haemophilus influenzae 2019.* Biochemistry, 2000. **39**(40): p. 12406-14.
57. Harvey, D.J., *Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates.* Mass Spectrom Rev, 1999. **18**(6): p. 349-450.
58. Domon, B. and C.E. Costello, *A systematic Nomenclature for Carbohydrate Fragmentations in FAB-MS/MS Spectra of Glycoconjugates.* Glycoconjugate, 1988. **5**: p. 397-409.
59. Loss, A., et al., *SWEET-DB: an attempt to create annotated data collections for carbohydrates.* Nucleic Acids Res, 2002. **30**(1): p. 405-8.
60. van Kuik, J.A., K. Hard, and J.F. Vliegenthart, *A 1H NMR database computer program for the analysis of the primary structure of complex carbohydrates.* Carbohydr Res, 1992. **235**: p. 53-68.
61. van Kuik, J.A. and J.F. Vliegenthart, *Databases of complex carbohydrates.* Trends Biotechnol, 1992. **10**(6): p. 182-5.
62. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
63. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence data bank and its supplement TrEMBL.* Nucleic Acids Res, 1997. **25**(1): p. 31-6.
64. O'Donovan, C., et al., *High-quality protein knowledge resource: SWISS-PROT and TrEMBL.* Brief Bioinform, 2002. **3**(3): p. 275-84.
65. *Editorial.* Nucleic Acids Res, 2003. **31**(14): p. 3869-3871.
66. Rindflesch, T.C., et al., *EDGAR: extraction of drugs, genes and relations from the biomedical literature.* Pac Symp Biocomput, 2000: p. 517-28.
67. Tanabe, L., et al., *MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling.* Biotechniques, 1999. **27**(6): p. 1210-4, 1216-7.
68. Bohne, A., E. Lang, and C. von der Lieth, *W3-SWEET: Carbohydrate Modeling by Internet.* J. Mol. Model., 1998. **4**: p. 33-43.
69. Bohne, A., E. Lang, and C.W. von der Lieth, *SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides.* Bioinformatics, 1999. **15**(9): p. 767-8.
70. Rötzer, F., *Speed matters I-IV.* Telepolis, 2000.
71. Marchal, I., et al., *Bioinformatics in Glycobiology.* Biochemie, 2003. **85**(1-2): p. 75-81.
72. Karas, M., et al., *Matrix-assisted ultraviolet laser desorption of non-volatile compounds.* Int J Mass Spectrom Ion Processes, 1987. **78**: p. 53-68.
73. Dell, A. and H.R. Morris, *Glycoprotein structure determination by mass spectrometry.* Science, 2001. **291**(5512): p. 2351-6.
74. Gillece-Castro, B.L. and A.L. Burlingame, *Oligosaccharide characterization with high-energy collision-induced dissociation mass spectrometry.* Methods Enzymol, 1990. **193**: p. 689-712.
75. Medzihradszky, K.F., et al., *Structure determination of O-linked glycopeptides by tandem mass spectrometry.* Biomed Environ Mass Spectrom, 1990. **19**(12): p. 777-81.

76. Teng-umnuay, P., et al., *The cytoplasmic F-box binding protein SKP1 contains a novel pentasaccharide linked to hydroxyproline in Dictyostelium.* J Biol Chem, 1998. **273**(29): p. 18242-9.
77. Metelmann, W., J. Peter-Katalinic, and J. Muthing, *Gangliosides from human granulocytes: a nano-ESI QTOF mass spectrometry fucosylation study of low abundance species in complex mixtures.* J Am Soc Mass Spectrom, 2001. **12**(8): p. 964-73.
78. Morris, H.R., et al., *Fast atom bombardment: a new mass spectrometric method for peptide sequence analysis.* Biochem Biophys Res Commun, 1981. **101**(2): p. 623-31.
79. Ii, T., et al., *Fast atom bombardment and electrospray ionization tandem mass spectrometry of sulfated Lewis(x) trisaccharides.* J Biochem (Tokyo), 1995. **118**(3): p. 526-33.
80. Huberty, M.C., et al., *Site-specific carbohydrate identification in recombinant proteins using MALD-TOF MS.* Anal Chem, 1993. **65**(20): p. 2791-800.
81. Ii, T., Y. Ohashi, and Y. Nagai, *Structural elucidation of underivatized gangliosides by electrospray-ionization tandem mass spectrometry (ESIMS/MS).* Carbohydr Res, 1995. **273**(1): p. 27-40.
82. Shen, X. and H. Perreault, *Characterization of carbohydrates using a combination of derivatization, high-performance liquid chromatography and mass spectrometry.* J Chromatogr A, 1998. **811**(1-2): p. 47-59.
83. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.
84. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
85. Wu, C.H., et al., *The Protein Information Resource: an integrated public resource of functional annotation of proteins.* Nucleic Acids Res, 2002. **30**(1): p. 35-7.
86. Cooper, C.A., E. Gasteiger, and N.H. Packer, *GlycoMod--a software tool for determining glycosylation compositions from mass spectrometric data.* Proteomics, 2001. **1**(2): p. 340-9.
87. Gaucher, S.P., J. Morrow, and J.A. Leary, *STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry.* Anal Chem, 2000. **72**(11): p. 2331-6.
88. Okamoto, M., et al., *High-sensitivity detection and postsource decay of 2-aminopyridine-derivatized oligosaccharides with matrix-assisted laser desorption/ionization mass spectrometry.* Anal Chem, 1997. **69**(15): p. 2919-26.
89. Harvey, D.J., *Electrospray mass spectrometry and fragmentation of N-linked carbohydrates derivatized at the reducing terminus.* J Am Soc Mass Spectrom, 2000. **11**(10): p. 900-15.
90. Hase, S., T. Ibuki, and T. Ikenaka, *Reexamination of the pyridylamination used for fluorescence labeling of oligosaccharides and its application to glycoproteins.* J Biochem (Tokyo), 1984. **95**(1): p. 197-203.
91. Okamoto, M., K. Takahashi, and T. Doi, *Sensitive detection and structural characterization of trimethyl(p-aminophenyl)-ammonium-derivatized oligosaccharides by electrospray ionization-mass spectrometry and tandem mass spectrometry.* Rapid Commun Mass Spectrom, 1995. **9**(8): p. 641-3.

92. Geyer, H., et al., *Core structures of polysialylated glycans present in neural cell adhesion molecule from newborn mouse brain.* Eur. J. Biochem., 2001. **268**: p. 6587-99.
93. Harvey, D.J., R.H. Bateman, and M.R. Green, *High-energy collision-induced fragmentation of complex oligosaccharides ionized by matrix-assisted laser desorption/ionization mass spectrometry.* J Mass Spectrom, 1997. **32**(2): p. 167-87.
94. Harvey, D.J., *Collision-induced fragmentation of underivatized N-linked carbohydrates ionized by electrospray.* J Mass Spectrom, 2000. **35**(10): p. 1178-90.
95. Harvey, D.J., et al., *"Internal residue loss": rearrangements occurring during the fragmentation of carbohydrates derivatized at the reducing terminus.* Anal Chem, 2002. **74**(4): p. 734-40.
96. Kurokawa, T., et al., *Hemocyanin from the keyhole limpet Megathura crenulata (KLH) carries a novel type of N-glycans with Gal(beta1-6)Man-motifs.* Eur J Biochem, 2002. **269**(22): p. 5459-73.
97. Royle, L., et al., *An analytical and structural database provides a strategy for sequencing O-glycans from microgram quantities of glycoproteins.* Anal Biochem, 2002. **304**(1): p. 70-90.
98. Kuster, B., et al., *Sequencing of N-linked oligosaccharides directly from protein gels: in-gel deglycosylation followed by matrix-assisted laser desorption/ionization mass spectrometry and normal-phase high-performance liquid chromatography.* Anal Biochem, 1997. **250**(1): p. 82-101.
99. Burks, C., et al., *The GenBank nucleic acid sequence database.* Comput Appl Biosci, 1985. **1**(4): p. 225-33.
100. Bucher, P. and A. Bairoch, *A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 53-61.
101. Falquet, L., et al., *The PROSITE database, its status in 2002.* Nucleic Acids Res, 2002. **30**(1): p. 235-8.
102. Sigrist, C.J., et al., *PROSITE: a documented database using patterns and profiles as motif descriptors.* Brief Bioinform, 2002. **3**(3): p. 265-74.
103. Bernstein, F.C., et al., *The Protein Data Bank. A computer-based archival file for macromolecular structures.* Eur J Biochem, 1977. **80**(2): p. 319-24.
104. Gupta, R., et al., *O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins.* Nucleic Acids Res, 1999. **27**(1): p. 370-2.
105. Hansen, J.E., et al., *O-GLYCBASE version 2.0: a revised database of O-glycosylated proteins.* Nucleic Acids Res, 1997. **25**(1): p. 278-82.
106. Hansen, J.E., et al., *Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase.* Biochem J, 1995. **308 (Pt 3)**: p. 801-13.
107. Hansen, J.E., et al., *NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility.* Glycoconj J, 1998. **15**(2): p. 115-30.
108. Cooper, C.A., et al., *GlycosuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources.* Nucleic Acids Research, 2001. **29**(1): p. 332-335.

109. Cooper, C.A., et al., *GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update.* Nucleic Acids Res, 2003. **31**(1): p. 511-3.
110. Doubet, S., et al., *The complex Carbohydrate Structure Database.* TIBS, 1989. **14**: p. 475-477.
111. Bohne-Lang, A., et al., *LINUCS: linear notation for unique description of carbohydrate sequences.* Carbohydr Res, 2001. **336**(1): p. 1-11.
112. Doubet, S. and P. Albersheim, *CarbBank.* Glycobiology, 1992. **2**(6): p. 505.
113. Gattiker, A., et al., *Automated annotation of microbial proteomes in SWISS-PROT.* Comput Biol Chem, 2003. **27**(1): p. 49-58.
114. Henry, B., et al., *NMR study of a Lewis(X) pentasaccharide derivative: solution structure and interaction with cations.* Carbohydr Res, 1999. **315**(1-2): p. 48-62.
115. Niedziela, T., et al., *Core oligosaccharides of Plesiomonas shigelloides O54:H2 (strain CNCTC 113/92): structural and serological analysis of the lipopolysaccharide core region, the O-antigen biological repeating unit, and the linkage between them.* J Biol Chem, 2002. **277**(14): p. 11653-63.
116. Soria-Diaz, M.E., et al., *Structural determination of the lipo-chitin oligosaccharide nodulation signals produced by Rhizobium giardinii bv. giardinii H152.* Carbohydr Res, 2003. **338**(3): p. 237-50.
117. Kuster, B., et al., *Glycosylation analysis of gel-separated proteins.* Proteomics, 2001. **1**(2): p. 350-61.
118. Pfenninger, A., et al., *Structural analysis of underivatized neutral human milk oligosaccharides in the negative ion mode by nano-electrospray MS(n) (part 2: application to isomeric mixtures).* J Am Soc Mass Spectrom, 2002. **13**(11): p. 1341-8.
119. Leonard, J.E., J.B. Colombe, and J.L. Levy, *Finding relevant references to genes and proteins in Medline using a Bayesian approach.* Bioinformatics, 2002. **18**(11): p. 1515-22.
120. de Bruijn, B. and J. Martin, *Getting to the (c)ore of knowledge: mining biomedical literature.* Int J Med Inf, 2002. **67**(1-3): p. 7-18.
121. Shah, P.K., et al., *Information extraction from full text scientific articles: Where are the keywords?* BMC Bioinformatics, 2003. **4**(1): p. 20.
122. Al-Sanae, H., et al., *Comparison of lactose intolerance in healthy Kuwaiti and Asian volunteers.* Med Princ Pract, 2003. **12**(3): p. 160-3.
123. Gonzalez, R., et al., *In vitro release of sodium diclofenac from a central core matrix tablet aimed for colonic drug delivery.* Eur J Pharm Sci, 2003. **20**(1): p. 125-131.
124. Ferrero, C., I. Bravo, and M.R. Jimenez-Castellanos, *Drug release kinetics and fronts movement studies from methyl methacrylate (MMA) copolymer matrix tablets: effect of copolymer type and matrix porosity.* J Control Release, 2003. **92**(1-2): p. 69-82.
125. Li, P., et al., *[Chemical constituents of Actinidia kolomikta (Rupr. et Maxim.) Planch.].* Zhongguo Zhong Yao Za Zhi, 1992. **17**(7): p. 420-1, 446.
126. Ray, E.T., *Einführung in XML*. 2001.
127. Eklund, R. and G. Widmalm, *Molecular dynamics simulations of an oligosaccharide using a force field modified for carbohydrates.* Carbohydr Res, 2003. **338**(5): p. 393-8.

128. Muldoon, J., et al., *Structure of an acidic polysaccharide from the marine bacterium Pseudoalteromonas flavipulchra NCIMB 2033(T).* Carbohydr Res, 2003. **338**(5): p. 459-62.
129. Li, M., X. Han, and B. Yu, *Synthesis of monomethylated dioscin derivatives and their antitumor activities.* Carbohydr Res, 2003. **338**(2): p. 117-21.
130. Dasgupta, S., et al., *Molecular characterization and immunohistochemical localization of IV(4)GalNAcGgOse(4)Cer: a naturally occurring novel neutral glycosphingolipid in bovine brain.* Glycobiology, 2000. **10**(1): p. 1-9.
131. Wuhrer, M., et al., *Schistosoma mansoni cercarial glycolipids are dominated by Lewis X and pseudo-Lewis Y structures.* Glycobiology, 2000. **10**(1): p. 89-101.
132. Muller-Loennies, S., et al., *Characterization of high affinity monoclonal antibodies specific for chlamydial lipopolysaccharide.* Glycobiology, 2000. **10**(2): p. 121-30.
133. Blaschke, C., L. Hirschman, and A. Valencia, *Information extraction in molecular biology.* Brief Bioinform, 2002. **3**(2): p. 154-65.
134. Nord, L.I. and L. Kenne, *Novel acetylated triterpenoid saponins in a chromatographic fraction from Quillaja saponaria Molina.* Carbohydr Res, 2000. **329**(4): p. 817-29.
135. Nord, L.I. and L. Kenne, *Separation and structural analysis of saponins in a bark extract from Quillaja saponaria Molina.* Carbohydr Res, 1999. **320**(1-2): p. 70-81.
136. Nord, L.I., L. Kenne, and S.P. Jacobsson, *Multivariate analysis of 1H NMR spectra for Saponins from Quillaja saponaria Molina.* Analytica Chimica Acta, 2001. **446**: p. 199-209.
137. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* J Comp Chem, 1998. **19**(14): p. 1639-1662.
138. Morris, G.M., et al., *Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4.* J Comput Aided Mol Des, 1996. **10**(4): p. 293-304.
139. Weiner, S.J., et al., *A new force field for molecular mechanical simulation of nucleic acids and proteins.* J Am Chem Soc, 1984. **106**: p. 765-784.
140. Hawkey, C.J., *COX-2 inhibitors.* Lancet, 1999. **353**(9149): p. 307-14.
141. Vane, J.R., *Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs.* Nat New Biol, 1971. **231**(25): p. 232-5.
142. O'Banion, M.K., et al., *A serum- and glucocorticoid-regulated 4-kilobase mRNA encodes a cyclooxygenase-related protein.* J Biol Chem, 1991. **266**(34): p. 23261-7.
143. Kujubu, D.A., et al., *TIS10, a phorbol ester tumor promoter-inducible mRNA from Swiss 3T3 cells, encodes a novel prostaglandin synthase/cyclooxygenase homologue.* J Biol Chem, 1991. **266**(20): p. 12866-72.
144. Bensen, W.G., *Antiinflammatory and analgesic efficacy of COX-2 specific inhibition: from investigational trials to clinical experience.* J Rheumatol Suppl, 2000. **60**: p. 17-24.
145. Reininger, E., *Vergleichende phytochemische und pharmakologische Untersuchungen zur Hemmung der Prostaglandin-H-Synthase Isoenzyme mit Arzneidrogen der chinesischen Medizin, Insbesondere Platycodi radix und*

Chaenomelis fructus, in *Mathematisch-Naturwissenschaftliche Fakultät*. 2001, Heinrich-Heine-Universität: Düsseldorf. p. 1-296.
146. Garcia-Nieto, R., C. Perez, and F. Gago, *Automated docking and molecular dynamics simulations of nimesulide in the cyclooxygenase active site of human prostaglandin-endoperoxide synthase-2 (COX-2).* J Comput Aided Mol Des, 2000. **14**(2): p. 147-60.
147. Selinsky, B.S., et al., *Structural analysis of NSAID binding by prostaglandin H2 synthase: time-dependent and time-independent inhibitors elicit identical enzyme conformations.* Biochemistry, 2001. **40**(17): p. 5172-80.
148. Kurumbail, R.G., et al., *Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents.* Nature, 1996. **384**(6610): p. 644-8.
149. Loll, P.J., D. Picot, and R.M. Garavito, *The structural basis of aspirin activity inferred from the crystal structure of inactivated prostaglandin H2 synthase.* Nat Struct Biol, 1995. **2**(8): p. 637-43.
150. Meade, E.A., W.L. Smith, and D.L. DeWitt, *Differential inhibition of prostaglandin endoperoxide synthase (cyclooxygenase) isozymes by aspirin and other non-steroidal anti-inflammatory drugs.* J Biol Chem, 1993. **268**(9): p. 6610-4.
151. Garavito, R.M. and D.L. DeWitt, *The cyclooxygenase isoforms: structural insights into the conversion of arachidonic acid to prostaglandins.* Biochim Biophys Acta, 1999. **1441**(2-3): p. 278-87.
152. Kalgutkar, A.S., et al., *Aspirin-like molecules that covalently inactivate cyclooxygenase-2.* Science, 1998. **280**(5367): p. 1268-70.
153. Paulus, K., *Untersuchungen zur Leukotrienbiosynthese-hemmenden Wirkung chinesischer Arzneidrogen, insbesondere von Salvia miltiorrhizae radix*, in *Mathematisch-Naturwissenschaftliche Fakultät*. 2002, Heinrich-Heine-Universität: Düsseldorf. p. 1-182.
154. Schwarte, A., *Phytochemische und pharmakologische Untersuchungen der Wurzel von Sophora flavescens, unter besonderer Berücksichtigung ihrer Wirkung auf die Leukotrien- und Prostaglandinbiosynthese*, in *Mathematisch-Naturwissenschaftliche Fakultät*. 2002, Heinrich-Heine-Universität: Düsseldorf. p. 1-240.
155. Seto, E., Y. Shi, and T. Shenk, *YY1 is an initiator sequence-binding protein that directs and activates transcription in vitro.* Nature, 1991. **354**(6350): p. 241-5.
156. Basu, A., et al., *Identification of a transcriptional initiator element in the cytochrome c oxidase subunit Vb promoter which binds to transcription factors NF-E1 (YY-1, delta) and Sp1.* J Biol Chem, 1993. **268**(6): p. 4188-96.
157. Usheva, A. and T. Shenk, *TATA-binding protein-independent initiation: YY1, TFIIB, and RNA polymerase II direct basal transcription on supercoiled template DNA.* Cell, 1994. **76**(6): p. 1115-21.
158. Houbaviy, H.B., et al., *Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator.* Proc Natl Acad Sci U S A, 1996. **93**(24): p. 13577-82.
159. Sadowski, J., M. Wagener, and J. Gasteiger, *CORINA: Automatic Generation of High-Quality 3D-Molecular Models for Application in QSAR*, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological*

Applications, F. Sanz, J. Giraldo, and F. Manaut, Editors. 1995, Prous Science Publishers. p. 646-651.

14 Danksagung

Ich möchte allen, die mir bei der Durchführung dieser Arbeit zur Seite gestanden haben, an dieser Stelle sehr herzlich danken.

Mein besonderer Dank gilt Herrn Dr. Claus-Wilhelm von der Lieth, für die großartige fachliche Unterstützung und seine unermüdliche Bereitschaft mir zu helfen.

Thomas Lütteke und Martin Frank danke ich für die Bereitstellung des *UMF-KONVERTERS*.

Andreas Bohne-Lang danke ich für die Programmierung von *SWEET-II*.

Alexander Loss danke ich für die Entwicklung der Schnittstelle zum Erstellen der LinucsID und der Pflege der *SWEET-DB*.

Johanna Lohmann danke ich für die Geduld und das häufige Korrekturlesen meiner Arbeit.

Ruth Lohmann danke ich für die Hilfe bei pharmazeutischen Fragestellungen.

Karin Bettinger danke ich für die angenehme Gesellschaft im Labor und für ihre Hilfe bei allen Fragestellungen.

Wolf D. Lehmann danke ich für die Unterstützung bei der Lösung von massenspektrometrischen Fragestellungen.

Bill Hull danke ich für die Möglichkeit in seiner Abteilung die Arbeit anzufertigen und für die Hilfe bei NMR-relevanten Fragestellungen

Franziska Beyer und Jan Kerschgens danke ich für die Hilfe und Anregungen bei der Entwicklung der Arbeitsumgebung zur Eingabe von NMR-Spektren.

Silke Templin, Anette Henke, Fadi Qa’dan und Mathias Hambörger danke ich für die Unterstützung während meines Pharmazie-Studiums.

Kerstin Paulus, Anne Schwarte, Birgit Dietz und Franka Teuscher danke ich für den angenehmen wissenschaftlichen Feedback aus Düsseldorf

15 Lebenslauf

von	Klaus Karl Lohmann	
geboren am:	27. August 1968 in Wiedenbrück	
Eltern:	Carl und Johanna Lohmann, geb. Seelhorst	
Familienstand:	ledig	
Staatsangehörigkeit:	deutsch	
Schulbildung:	1975-1979:	Brinkmannschule Langenberg
	1979-1988:	Marienschule Lippstadt
Wehrdienst:	1988-1989:	Ableistung des Wehrdienstes bei der 3./PzBtl. 513
Studium:	1989-1991:	Informatikstudium an der Gesamthochschule Kaiserslautern
	1991-1995:	Pharmaziestudium an der Westfälischen-Wilhelms-Universität Münster
	30.11.1995:	2. Staatsexamen
Praktisches Jahr:	12/1995-11/96:	Pharmaziepraktikum in der Abtei-Apotheke, Liesborn
Abschluß:	12.12.1996:	3. Staatsexamen
	07.01.1997:	Approbation als Apotheker
Berufsausübung:	02/1997-03/1999:	Vertretungstätigkeit im Kammerbezirk Westfalen-Lippe
	04/1999-04/2001:	Systembetreuer am Institut für Pharmazeutische Biologie der Heinrich-Heine-Universität Düsseldorf
	seit 05/2001:	Doktorand am Deutschen Krebsforschungszentrum Heidelberg