Linux, Knoppix, Mac OS X, Open Source: Vorteile von Unix et al. in Chemie & Biologie

Teil 8: Computational Chemistry

Röbbe Wünschiers

Molekülstrukturen zeichnen? Kristallstrukturen anzeigen? Moleküldynamiken simulieren? NMR- oder IR-Spektren vorhersagen? Das alles und noch mehr ist möglich mit Vigyaan Linux. In der vergangenen Woche habe Sie die graphische Oberfläche von Linux näher kennen gelernt. In dieser Woche möchte ich Ihnen einige interessante Programme für Chemiker und Biologen vorstellen.

Chemoinformatik und Bioinformatik sind zwei Forschungsgebiete die in besonderem Maße auf die Anwendung von Computern zur Lösung von Problemen angewiesen sind. Allerdings sind die Begriffe Chemo- und Bioinformatik vage. Sie können alles von der puren mathematischen Algorithmik bis hin zur Anwendung eines Tabellenkalkulationsprogamms zur Berechnung von Puffergleichgewichten beinhalten. Im Englischen wird zwischen *Chemo-/Bioinformatics* und *Computational Chemistry/Biology* unterschieden. Letzteres meint die angewandte Seite: der Computer wird eingesetzt um Probleme in den Bereichen Chemie und Biologie zu lösen. *Chemo-* und *Bioinformatics* sind dagegen eigenständige Forschungsgebiete. Leider gibt es für *Computational Chemistry* oder *Biology* keine gängigen Übersetzungen. Computergestützte Chemie? Das hört sich komisch an und fließt nicht locker über die Lippen. So haben sich die Begriffe Chemoinformatik und Bioinformatik eingebürgert, bestenfalls noch mit der Einschränkung Angewandte Chemo- und Bioinformatik. Mit diesem Thema wollen wir uns nun beschäftigen.

Vigyaan Knoppix

Des öfteren war bereits von Knoppix, einer freien Linuxdistribution die von CD-ROM startet, die Rede. Der Chemiker Pratul Agarwal hat eine Knoppixdistribution zusammengestellt, die wesentliche Programme für die Bereiche Chemie und Biologie enthält. Diese Distribution namens Vigyaan (der Name kommt aus dem Sanskrit und bedeutet soviel wie Wissenschaft) ist kostenlos unter http://www.vigyaancd.org herunterzuladen. Wenn Sie über keine oder nur eine langsame Internetverbindung verfügen, dann können Sie die Vigyaan CD für 5 Euro inklusive Porto und Verpackung bei der CLB-Redaktion (redaktion@clb.de, Stichwort Vigyaan) bestellen. Im ersten Teil dieser Serie (CLB 11/2003) wurde die Verwendung von Knoppix und somit auch von Vigyaan kurz besprochen. Vigyaan-Knoppix enthält eine Vielzahl von Programmen für Chemiker und Biologen. Eine Übersicht über die Arbeitsfläche (*Desktop*) von Vigyaan-Knoppix ist in Abbildung 1 zu sehen. Um Ihnen einen Überblick über die Möglichkeiten zu geben, seien die wichtigsten Applikationen kurz vorgestellt.

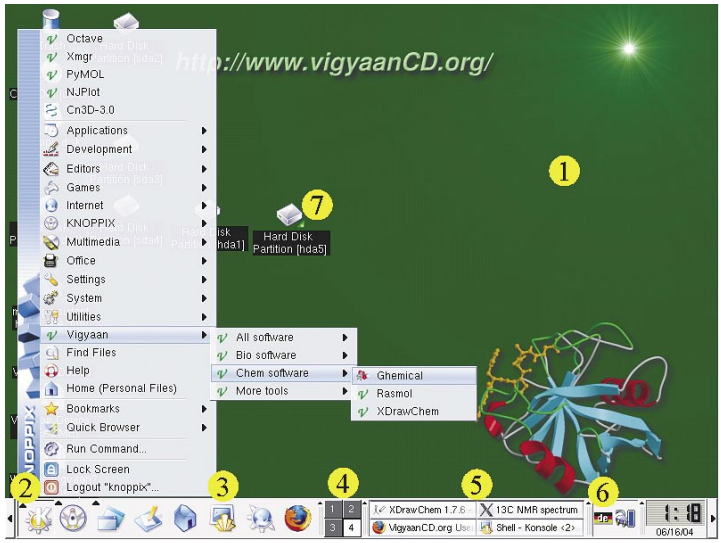
3D Molekülstruktur-Betrachter

Eine der ersten chemoinformatischen Fragestellung war die Rekonstruktion einer Molekülstruktur aus ihren spektroskopischen Eigenschaften (insbesondere der Röntgenbeugung). Es ist also nicht verwunderlich, das es eine Vielzahl von Programmen zur Darstellung dreidimensionaler Molekülstrukturen gibt (sogenannte *Structure-Viewer*).

Rasmol – http://www.umass.edu/microbio/rasmol/index2.htm – Der Klassiker unter den Molekülstruktur-Viewern (siehe Abbildung 6). Rasmol ist für jede Computerplattform verfügbar und erfreut sich trotz seines hohen Alters großer Beliebtheit. Das Programm verfügt über eine eigene Skriptsprache, die komplexe Darstellungen von räumlichen molekularen Strukturen erlaubt (siehe Skript *sulphur.ras* weiter unten). Im zweiten Anwendungsbeispiel weiter unten werden wir Rasmol einsetzten.

JMol – http://jmol.sourceforge.net – Wie Rasmol, so dient auch JMol dem Betrachten von Molekülstruktu-

Abbildung 1: Vigyaan-Knoppix. (1) Arbeitsplatz (*Desktop*); (2) Startmenü; (3) Link zur Konsole (*Terminal, Shell*); (4) hier kann zwischen 4 Desktops gewechselt werden; (5) hier kann zwischen den geöffneten Programmen gewechselt werden; (6) ein Klick auf Fahne wechselt das Tastaturlayout; (7) die Festplatte wird erkannt und eingebunden.



ren. JMol ist etwas bedienerfreundlicher, dafür aber nicht so leistungsfähig.

PyMOL – http://pymol.sourceforge.net – PyMol ist einer der leistungsfähigsten freien Structure-Viewer. Allerdings benötigt PyMol erheblich mehr Rechenleistung als Rasmol.

Cn3D – http://www.ncbi.nih.gov/Structure/CN3D/cn3d.shtml – Cn3D kann als Plugin in den Internetbrowser eingebunden werden. Neben der Struktur kann in einem separaten Fenster auch die Proteinsequenz angezeigt werden.

Garlic – http://garlic.mefos.hr/garlic – Da die Homepage von Garlic nicht gut zugänglich ist, hier ein Link zu einer PDF-Version (*portable document format*; lesbar mit dem freien Adobe Acrobat Reader) der Internetseite: http://www-vis.lbl.gov/NERSC/Software/garlic/docs/pdf/garlic_website.pdf. Garlic ist ein Programm das auf die Visualisierung von Membranproteinen spezialisiert ist. Als Dateiformat dient das PDB Format.

2D Molekülstruktur Editor und mehr

Jedem Chemiker begegnet irgendwann die Aufgabe Strukturformeln zu zeichnen. Vigyaan-Knoppix bietet zu diesem Zweck nur ein Programm an.

XDrawChem – http://xdrawchem.sourceforge.net – Neben dem Zeichnen von Strukturformeln bietet XDrawChem unter anderem die Möglichkeit Infrarot- und ^{1}H beziehungsweise ^{13}C NMR-Spektren zu simulieren (siehe Abbildungen 2 und 3).

Abbildung 2: XDrawChem. Erstellte Strukturformel und berechnete Summenformel und Molekulargewicht des Tetrapeptids Cystein-Glycin-Prolin-Cystein (CGPC).

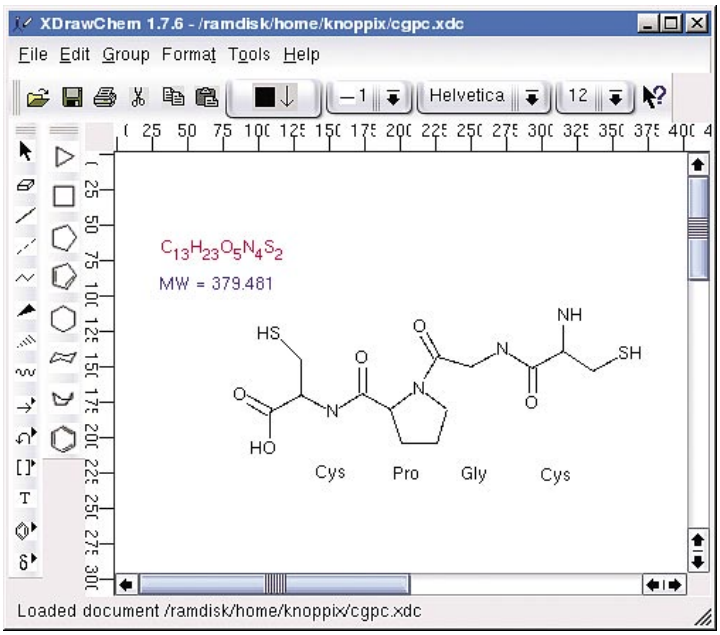
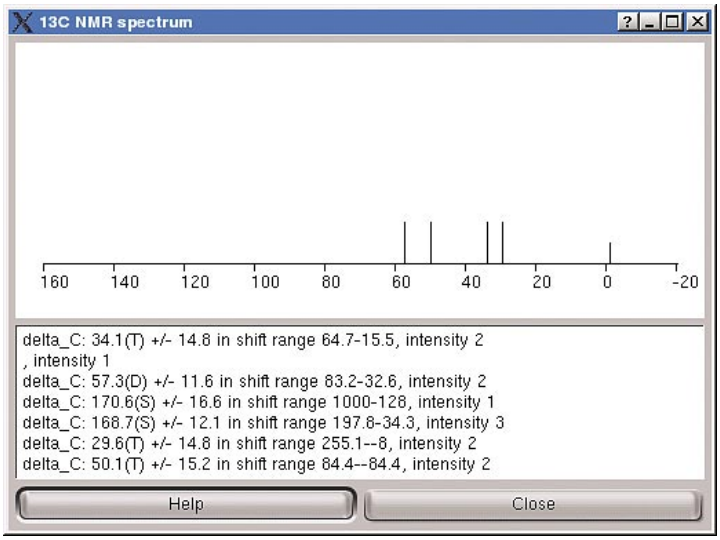


Abbildung 3: XDrawChem. Vorhergesagtes ^{13}C NMR-Spektrum der in Abbildung 2 gezeigten Struktur des Tetrapeptids CGPC.



Moleküldynamik & Quantenmechanik

Im Zeitalter von Internet und computergestützten Präsentationen vergisst man leicht, dass Computer eigentlich dazu entwickelt wurden um Rechenaufgaben zu lösen. Wenn Sie Ihren Prozessor mal richtig aufheizen möchten, dann füttern Sie ihn zum Beispiel mit Problemen aus der Quantenchemie oder Molekülmechanik ...

Ghemical – http://www.uku.fi/~thassine/ghemical – Ghemical bietet über eine graphische Benutzeroberfläche Zugang zu Programmen für komplexe chemoinformatische Probleme (siehe Abbildungen 4 und 5). Es ermöglicht unter anderem die Simulation von Moleküldynamiken sowie die Berechnung von Energiepotentialen und quantenmechanischen Zuständen von Strukturen. Strukturdaten können aus verschiedenen Dateiformaten importiert und exportiert werden. Im ersten Anwendungsbeispiel weiter unten werden wir Ghemical verwenden.

MPQC – http://aros.ca.sandia.gov/~cljanss/mpqc/index.html – MPQC (*massively parallel quantum chemistry*) ist ein Kommandozeilen Programm für quantenmechanische Berechnungen. Ghemical bietet zu einigen Teilprogrammen von MPQC einen graphischen Zugang.

Gromacs – http://www.gromacs.org – Gromacs dient der Berechnung von molekularen Dynamiken, wie beispielsweise der Simulation des Verhaltens von Molekülen in Lösungsmitteln.

PSI3 – http://www.psicode.org – Ein weiteres Programm für quantenmechanische Berechnungen.

DNA- & Proteinsequenz Analyse

Eine große Rolle spielt der Computer bei der Analyse von DNA- und Proteinsequenzen um insbesondere funktionale oder evolutive Zusammenhänge zu erkennen oder Sequenzmotive zu identifizieren. Für diese Aufgaben gibt es eine vergleichsweise große Auswahl an Programmen.

Emboss – http://www.hgmp.mrc.ac.uk/Software/EMBOSS – Unter dem Namen Emboss sind mehr als 100 Programme rund um die Verarbeitung von DNA- und Protein-Sequenzdaten vereint. Dazu gehören Programme zur Erstellung von Sequenzalignments, zur Identifizierung von Sequenzmotiven, der Genomanalyse und vieles mehr.

Artemis – http://www.sanger.ac.uk/Software/Artemis – Artemis ist ein leistungsfähiger DNA-Sequenz Betrachter. Das Program liest verschiedene Eingabeformate wie EMBL, Genbank oder Fasta. Die DNA-Sequenz nebst annotation wird angezeigt und ist editierbar.

ClustalX – http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html – ClustalX ist die graphische Benutzeroberfläche zu ClustalW, dem Klassiker unter den Sequenzalignment Programmen. Sowohl DNA- als auch Proteinsequenzen können unter Berücksichtigung verschiedender evolutionärer Modelle alignt werden. Zusätzlich werden die berechneten Distanzdaten abgespeichert. Sie können zum Beispiel mittels TreeView oder NJPlot zur Erstellung eines phylogentischen Baumes verwendet werden.

NCBI Toolkit – http://www.ncbi.nlm.nih.gov/BLAST – Das NCBI Toolkit vom amerikanischen *National Center for Biotechnology Information* bietet verschiedene kleinere Kommandozeilen orientierte Programme die mit den unterschiedlichen NCBI Datenbanken zusammenarbeiten. Am prominentesten ist wahrscheinlich Blast (*Basic Local Alignment Search Tool*). Das NCBI Toolkit erlaubt das Erstellen lokaler Sequenzdatenbanken.

Smile – http://www-igm.univ-mlv.fr/~marsan/smile_english.html – Smile ist ein Kommandozeilen orientiertes Programm das nach Sequenzmustern sucht. Es wurde ursprünglich entwickelt um Promotersequenzen (Genregulationseinheiten) in DNA-Sequenzen zu identifizieren. Smile kann aber ebenso auf Proteinsequenzen losgelassen werden. Die Basis zur Datenanalyse ist ein Suffixtree der von Smile intern erstellt wird.

Sonstige Tools

Schließlich seien einige Programme der Vigyaan-Knoppix Distribution erwähnt, die insbesondere für die Datenprozessierung und -visualisierung eine wichtige Rolle spielen.

Open Babel – http://openbabel.sourceforge.net – Open Babel dient der Konvertierung unterschiedlicher Dateiformate die in der Computational Chemistry Anwendung finden.

Bioperl – http://www.bioperl.org – Freunde der Programmiersprache Perl haben mit Bioperl Zugang zu einer umfangreichen Bibliothek von Modulen, welche die Bearbeitung biologischer Sequenzinformationen erlaubt.

GNU R – http://www.r-project.org – R ist eine freie Skriptsprache für die statistische Datenanalyse (die kommerzielle Version heißt S). Mit R stehen eine Vielzahl statistischer Modelle zur Verfügung. Darüber hinaus können die Daten graphisch dargestellt werden.

The Gimp – http://www.gimp.org – Gimp ist das frei zugängliche Pendant zu Adobe Photoshop. Mit Gimp können Sie beispielsweise mit Rasmol erstellte Bilddateien editieren und in andere Dateiformate konvertieren.

Xmgrace – http://plasma-gate.weizmann.ac.il/Xmgr – Xmgrace bietet die Möglichkeit Daten in xy- oder xyz-Plots graphisch darzustellen.

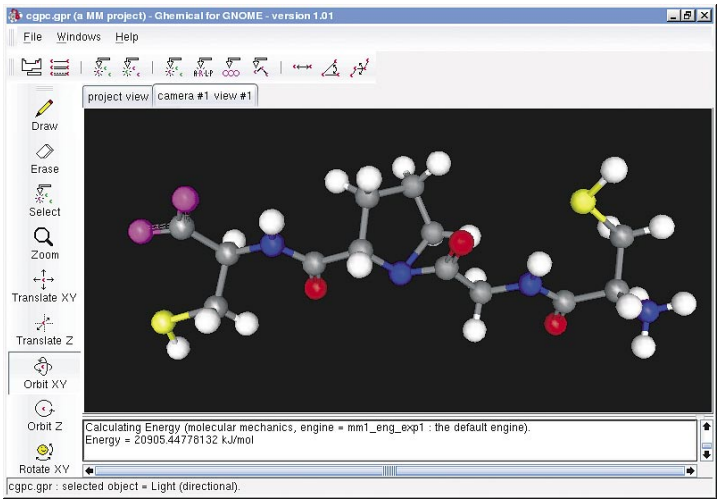
Beispiele

Es ist annähernd unmöglich alle erwähnten Softwarepakete vorzustellen. Selbst die detaillierte Vorstellung eines Programms würde den Rahmen dieser Kolumne sprengen. Stattdessen möchte ich zwei kleine Beispiele unter Verwendung von Ghemical und Rasmol vorstellen.

Faltung eines Tetrapeptids

In der Abbildung 2 ist die mit XDrawChem erstellte Strukturformeln des Tetrapeptids CGPC (Cystein-Glycin-Prolin-Cystein) gezeigt. Sowohl das Molekulargewicht wie auch die Summenformel wurden aus der

Abbildung 4: Ghemical. Native lineare Struktur des Tetrapeptids CGPC. Die berechnete Energie der Struktur beträgt rund 20905 kJ/mol.



erstellten Struktur berechnet. Abbildung 3 zeigt das vorhergesagte ^{13}C NMR-Spektrum des Tetrapeptids. Mit Hilfe von Ghemical wollen wir nun eine mögliche dreidimensionale Faltung des Tetrapeptids vorhersagen. Dazu müssen wir zunächst Ghemical im Startmenü öffnen (Abbildung 1). Wir beginnen nun ein neues Projekt durch klicken auf *File → New Project → ... Molecular Mechanics (.gpr)*. Nun klicken wir mit der rechten Maustaste in die schwarze Arbeitsfläche – es erscheint ein Menü. Wir wählen nun *Build → Sequence Builder (amino)* .... Es öffnet sich ein Fenster, welches wir durch einen Mausklick auf *OK* schließen. In dem nun offenen Fenster namens *Command Interpreter* ersetzen wir *AAA* durch die Aminosäuresequenz unseres Tetrapeptids: CGPC und klicken *OK*. In der Arbeitsfläche erscheint nun das Molekül. Wir öffnen wieder das Menü indem wir mit der rechten Maustaste auf die Arbeitsfläche klicken und wählen jetzt *Build → Hydrogens → Add*. Jetzt ist unsere Struktur vollständig. Blau erscheinen die Stickstoffatome, schwarz Kohlenstoff, weiß Wasserstoff, rot Sauerstoff und gelb Schwefel (Abbildung 4). Wenn Sie in der linken Menüleiste *Orbit XY* aktivieren, dann können Sie das Molekül drehen indem Sie mit der linken Maustaste in die Arbeitsfläche klicken und die Maustaste gedrückt halten während Sie die Maus bewegen. Lassen Sie uns nun die Energie dieser Struktur berechnen, indem wir, wie gehabt, das Menü öffnen und *Compute → Energy* auswählen. Das Ergebnis erscheint in der Statuszeile: *20905 kJ/mol*.

Abbildung 5: Ghemical. Von 100 zufälligen Konformationen des Tetrapeptids CGPC hat die hier gezeigte mit rund 38 kJ/mol die geringste Energie. Dies ist somit eine wahrscheinliche reale Konformation.

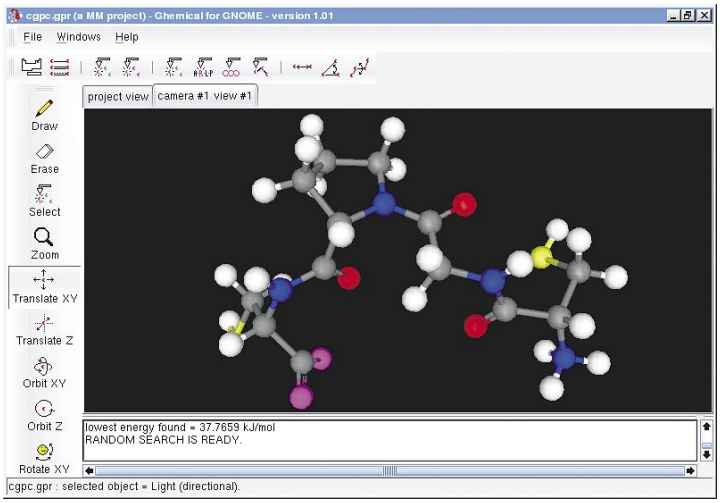
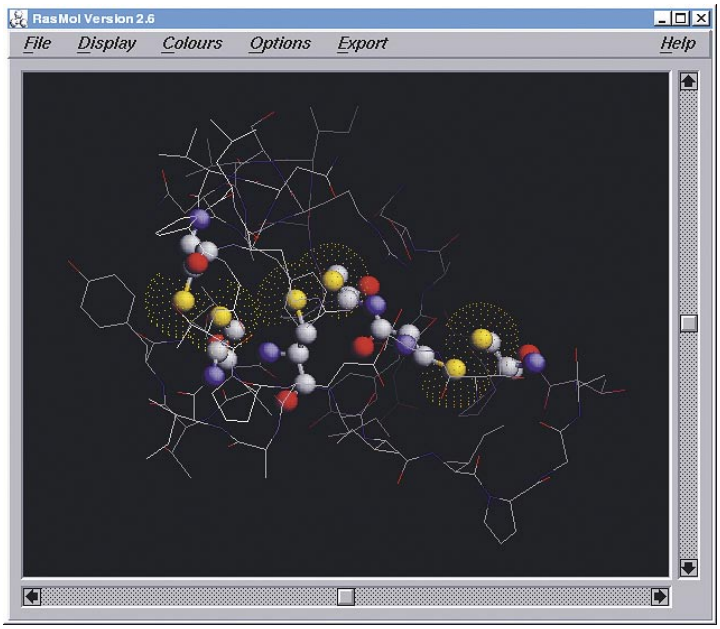


Abbildung 6: Rasmol. Darstellung der Struktur des Crambin Proteins. Die Cysteine sind hervorgehoben. Von allen Schwefelatomen sind die Van der Waals Radien gezeigt.



Jetzt lassen wir Ghemical nach dem Zufallsprinzip einen energiearmen stabilen Zustand suchen: klicken Sie den Menüpunkt *Compute → Random Conformational Search*. Es erscheint ein Fenster mit dem Inhalt *random_search 100 250*. Wir übernehmen diese Voreinstellung und klicken lediglich auf OK. Ausgehend von 100 zufälligen Ausgangskonformationen werden in 250 Iterationen Strukturkonformationen simuliert und deren Energie berechnet – dies kann einige Minuten dauern. Auf der Arbeitsfläche werden während der Berechnungen die aktuellen Konformationen angezeigt. Schließlich wird die energieärmest und damit wahrscheinlichste Konformation angezeigt (Abbildung 5). Die Energie wird in der Statuszeile angezeigt: *37.8 kJ/mol*. Erhalten Sie einen anderen Wert? Dann starten Sie die Berechnung von neuem.

Näheres Betrachten der resultierenden Struktur zeigt eine U-förmige Kurvatur des Peptidrückgrades. Dies entspricht der Beobachtung in Proteinen, dass die Aminosäure Prolin häufig Schleifen einleitet. Eine Dokumentation zu Ghemical können Sie mit dem Befehl `konqueror /usr/share/doc/ghemical/index.html` aus einer geöffneten Konsole laden. Eine Konsole (*Shell, Terminal*) wird geöffnet, wenn Sie in der unteren Symbolleiste auf Ihrem Bildschirm auf den Monitor mit der Muschel klicken (Abbildung 1).

Cysteine, Schwefel und Disulfidbrücken

Durch ihre Fähigkeit Disulfidbrücken auszubilden nehmen Cysteine in Proteinen eine Sonderstellung ein. Intermolekulare Disulfidbrücken bewirken eine Stabilisierung der Proteinstruktur. Die Trennung einer Disulfidbrücke durch Reduktion kann zu einer Konformationsänderung und somit zu eine Aktivitätsänderung bei Enzymen führen. Cysteine können daher auch an der Regulation von Enzymaktivitäten beteiligt sein. In Disulfidbrücken sind die Schwefelatome der

Cysteine weniger als 3 Ångström voneinander entfernt. In einer späteren Ausgabe werden wir ein Programm schreiben, um solche Schwefelatome aus einer Proteinstrukturdatei im PDB-Format herauszulesen. In diesem Beispiel sollen die Van der Waals Radien der Schwefelatome von Cysteinen in einer Proteinstruktur mittels Rasmol visualisiert werden. Als Struktur wählen wir Crambin, ein 46 Aminosäuren großes Protein aus Salatsamen. Geben Sie zunächst das nachfolgende Skript mit Hilfe des Texteditors **vim** ein. Wie bereits weiter oben beschrieben öffnen Sie dazu eine Konsole (Abbildung 1). Jetzt haben Sie eine Shell, wie Sie sie bereits kennen gelernt haben. Die Verwendung des Texteditors wurde bereits in Teil 4 (CLB 02/2004) besprochen.

Skript *sulphur.ras*

```
1      # Rasmol Skript
2      # von RW für CLB
3      # Lade Proteinstruktur
4      load pdb „~/Vigyaan/TINKER/
crambin.pdb“
5      # Zoom Struktur
6      zoom 151
7      # Waehle Cysteine
8      select cys
9      # Formatiere Struktur
10     wireframe 40
11     spacefill 150
12     # Waehle Schwefel
13     select sulphur
14     # VanDerWaals
15     dots
```

Nach Eingabe des Skriptes in eine Datei namens *sulphur.ras*, können Sie Rasmol mit dem Befehl **rasmol -script sulphur.ras** öffnen. Sie müssen den Befehl von jenem Verzeichnis aus ausführen, in welchem Sie die Skriptdatei gespeichert haben.Es sollte sich ein Fenster gemäß Abbildung 6 öffnen. Sie sehen als Punktwolken die Van der Waals Radien um die Schwefelatome der hervorgehobenen Cysteine des Crambin Proteins. Bewegen Sie den Mauszeiger auf der Struktur mit gedrückter linker Maustaste, so können Sie die Struktur drehen. Wechseln Sie zurück in die Konsole und geben das Kommando **quit** ein um Rasmol zu beenden. Eine Dokumentation zu Rasmol können Sie mit dem Befehl **konqueror /usr/share/doc/rasmol/rasmol.html** aus einer geöffneten Konsole laden.

Die beiden vorangegangenen Beispiele sollten Ihnen einen Einblick in die Möglichkeiten im Bereich von Computational Chemistry mit Linux bieten. Probieren Sie einfach verschiedene Dinge aus um Erfahrungen mit Vigyaan-Knoppix zu sammeln. Auf dem Desktop finden Sie auch einen Link zu verschieden Demos die Sie durcharbeiten können. Linux bietet kostenfrei leistungsfähige Programme!

Atomgewichte				Atommassen bezogen auf 12 C = 12,0000
Ag 107,868	Cs 132,9054	K 39,098	Pb 207,19	Sr 87,62
Al 26,981	Cu 63,546	Kr 83,80	Pd 106,4	Ta 180,9479
Ar 39,948	Dy 162,50	La 138,9155	Pr 140,9077	Tb 158,9254
As 74,9216	Er 167,26	Li 6,941	Pt 195,09	Tc 98,9062
Au 196,9665	Eu 151,96	Lu 174,97	Pu 244	Te 127,60
B 10,81	F 18,9984	Mg 24,305	Ra 226,02	Th 232,038
Ba 137,34	Fe 55,847	Mn 54,9380	Rb 85,467	Ti 47,90
Be 9,01218	Ga 69,72	Mo 95,94	Re 186,2	Tl 204,37
Bi 208,9804	Ge 72,59	N 14,0067	Rh 102,9055	Tm 168,9342
B 79,904	H 1,0079	Na 22,9898	Ru 101,07	U 238,029
C 12,011	He 4,0026	Nb 92,9064	S 32,064	V 50,944
Ca 40,08	Hf 178,49	Nd 144,24	Sb 121,75	W 183,85
Cd 112,40	Hg 200,59	Ne 20,179	Sc 44,9559	Xe 131,30
Ce 140,12	Ho 164,9304	Ni 58,71	Se 78,96	Y 88,909
Cl 35,453	I 126,9045	O 15,9994	Si 28,086	Yb 173,04
Co 58,9332	In 114,82	Os 190,2	Sm 150,35	Zn 65,38
Cr 51,996	Ir 192,22	P 30,9738	Sn 118,69	Zr 91,22