Problemseminar Bio-Datenbanken WS 2002/2003

Proteindatenbanken

Bearbeiter: Daniela Faust
Betreuer: Dr. Dieter Sosna
Prof. Erhard Rahm

Inhaltsverzeichnis

1. Einleitung

„ Die Forschung in Biologie und Chemie ist ungeheuer spannend, da die meisten Fragen nur oberflächlich angekratzt sind, wir aber heute über das Handwerkszeug verfügen, biologische Erscheinungen in ihrer molekularen Ursache zu verstehen – das ist doch eine phantastische Zeit! Zudem haben biologisch relevante Fragestellungen den großen Reiz der möglichen Anwendung. Man untersucht, und sieht beispielsweise durch Strukturanalyse zum erstenmal ein Molekül, mit dem man vielleicht ein neues Medikament zum Segen der Menschheit schaffen kann.“
Aus einem Interview [Cam98]
Robert Huber (Max-Planck-Institut für Biochemie bei München, 1988 mit H. Michel und Johann Deisenhofer, erhielt den Nobelpreis für Chemie für die Aufklärung der Struktur des photosynthetischen Reaktionszentrums des Purpurbakteriums Rhodopseudomonas viridis*)*

Letztlich ist eine Zelle, was sie ist, aufgrund ihrer Proteine.
Im Vordergrund steht die molekulare Struktur (Tertiär-/Quartärstruktur) von Proteinen und ihre derzeitige Analyse und Umsetzung in Datenbanken. Um diese zu verstehen wird auch kurz auf die Grundbausteine (Primär- und Sekundärstruktur) der Proteine eingegangen. Insgesamt wird ein Überblick gegeben über Proteinforschung, d. h. Bedeutung der Proteine, ihre Analyse und mögliche Umsetzungen der Analysen in Datenbanken und Vorstellung einiger Datenbanken.

2. Warum Proteinforschung?
2.1 Ein Beispiel

Man möchte mithelfen, ein Herbizid zu finden, das unschädlich ist für die Nutzpflanze, aber Unkraut vernichtet. Man geht so vor, dass die Struktur des Enzyms einer Nutzpflanze, Mais zum Beispiel, untersucht wird, um herauszufinden, welches Herbizid besonders schnell von diesem Maisenzym entgiftet wird. Nun macht man eigentlich das Gegenteil der üblichen Vorgehensweise. Üblicherweise „designed“ man Substanzen, die ein bestimmtes Enzym hemmen. Hier werden Herbizide vorgeschlagen, die von der Nutzpflanze, aber nicht vom Unkraut besonders schnell umgesetzt werden. Man analysiert die Struktur des Rezeptors und klärt den Funktionsmechanismus. Auf dieser Grundlage werden dann möglich weitere Herbizide geplant.[Cam98]

2.2 Die Funktion von Proteinen

Die Bedeutung der Proteine ist schon in ihrem Namen enthalten, der sich von dem griechischen Wort *proteios* für „erstrangig“ ableitet. Proteine machen mehr als 50 Prozent des Trockengewichts der meisten Zellen aus und dienen als Werkzeuge für fast alle Aktivitäten des Organismus. So werden sie als Elemente von Stützstrukturen, zur Speicherung und zum Transport anderer Stoffe, zur Signalübermittlung innerhalb des Organismus, zur Bewegung und zur Abwehr von Fremdstoffen eingesetzt (Tabelle1). Zusätzlich regulieren Proteine in Form von Enzymen den Stoffwechsel, indem sie chemische Reaktionen in der Zelle selektiv beschleunigen. Allgemein gesprochen sind Proteine darauf spezialisiert, andere Moleküle spezifisch und reversibel zu binden. Ein Mensch besitzt Zehntausende verschiedene Arten von Proteinen, jedes mit einer spezifischen Struktur und Funktion.
Proteine sind die strukturell am höchsten entwickelten Moleküle, die wir kennen. In Übereinstimmung mit ihren verschiedenartigen Funktionen variieren sie auch stark in ihrer Struktur: Jeder Proteintyp besitzt eine einzigartige dreidimensionale Gestalt oder

Konformation. Trotz ihrer Verschiedenartigkeit sind alle Proteine Polymere, die aus demselben Satz von 20 Aminosäuren aufgebaut sind, den universellen Monomeren der Proteine. Polymere aus Aminosäuren werden Polypeptide genannt. Ein Protein besteht aus einer oder mehreren Polypeptidketten, die in spezifischer Konformation gefaltet und gewunden sind.

Tabelle1: Ein Überblick über die Proteinfunktionen [Cam98]

Proteintyp	**Funktion**
Strukturproteine	Halt
Speicherproteine	Speicherung von Aminosäuren
Transportproteine	Transport von Stoffen
Hormonelle Proteine	Koordination der Aktivitäten eines Organismus
Rezeptorproteine	Reaktion einer Zelle auf chemische Reize
Kontraktile Proteine	Bewegung
Abwehrproteine	Schutz vor Krankheit
Enzymatische Proteine	selektive Beschleunigung chemischer Reaktionen

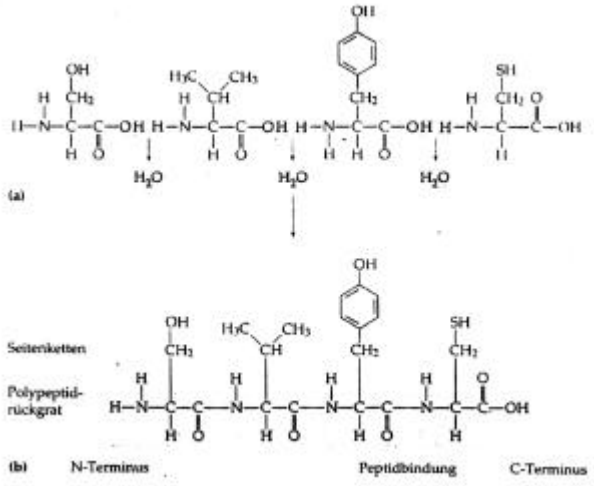
Ein funktionsfähiges Protein ist nicht *bloß* eine Polypeptidkette, sondern besteht aus einem oder mehreren Polypeptiden, die präzise zu einem Molekül von einzigartiger Gestalt gewunden, gefaltet und gedreht sind. Ein Polypeptid enthält Information in Form seiner Aminosäuresequenz, und genau diese Information bestimmt, welche dreidimensionale Konformation das Protein annimmt. Viel Proteine sind globulär, während andere eine langgestreckte Form haben. Innerhalb dieser Kategorien gibt es unzählige mögliche Variationen.
Die spezifische Konformation eines Proteins bestimmt seine biologische Wirkung. Fast in jedem Fall hängt die Funktion eines Proteins von seiner Fähigkeit ab, ein anderes Molekül zu erkennen und zu binden.

2.3 Die Struktur von Proteinen

Siehe auch Einführungsvortrag „Einführung in die Bioinformatik und die Bio-Datenbanken“.
Die Primärstruktur
ist die einzigartige Abfolge seiner Aminosäuren.

Abb. I) Polpeptidketten [Cam98]

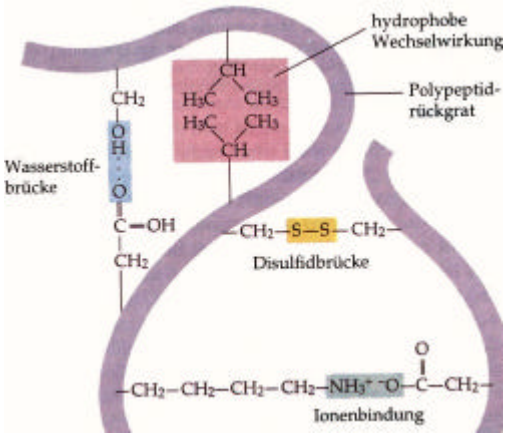
a) In Peptidbindungen, die durch Kondensationsreaktionen entstehen, wird die Carboxylgruppe einer Aminosäure mit der Aminogruppe der nächsten verbunden.
b) Das Polypeptid besitzt ein sich wiederholendes Rückgrat, an das die Aminosäureseitenketten geknüpft sind.



Die Sekundärstruktur
entsteht durch Windungen und Faltungen resultierend aus Wasserstoffbrücken, die in regelmäßigen Abständen entlang des Polypeptidrückgrates auftreten. → alpha-Helix, beta-Faltblatt

Die Tertiärstruktur
besteht aus unregelmäßigen Windungen, welche durch chemische Bindungen zwischen den Seitenketten der verschiedenen Aminosäuren stabilisiert werden.

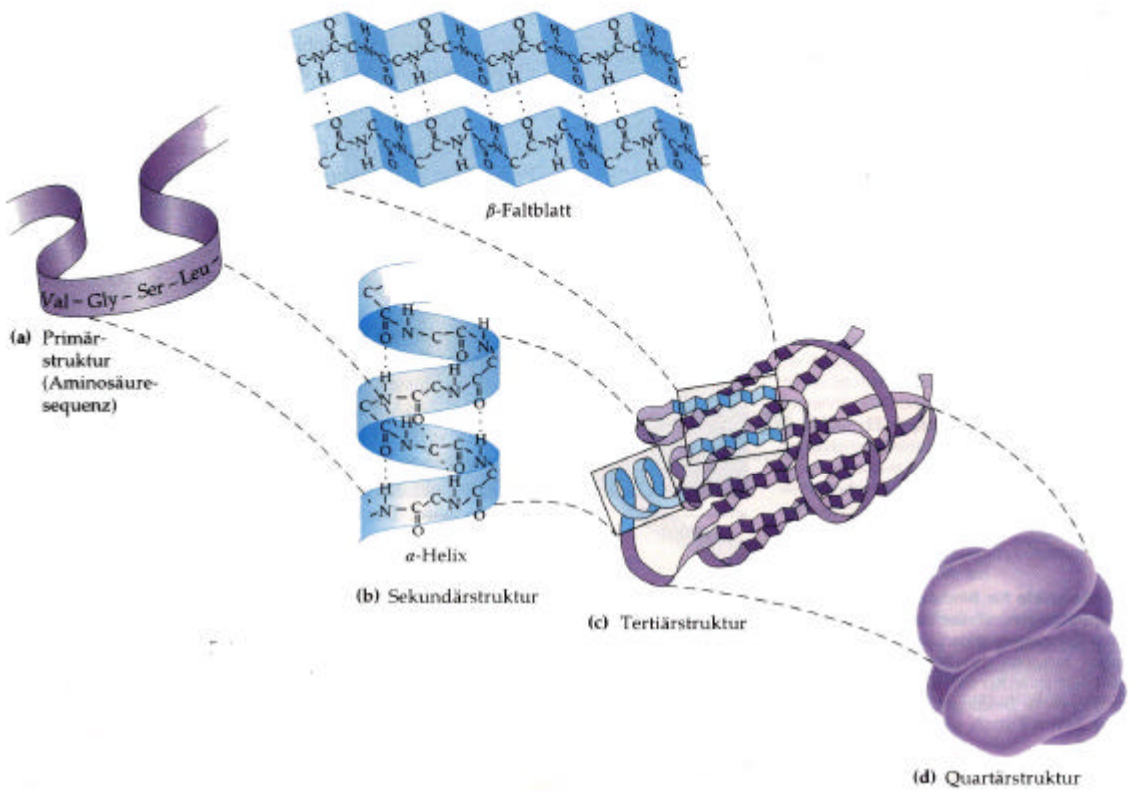
Abb. II) Beispiele für Bindungen, die zur Tertiärstruktur eines Proteins beitragen [Cam98]
Wasserstoffbrücken, Ionenbindungen und hydrophobe Wechselwirkungen (van-der-Waals-Kräfte) sind schwache Bindungen zwischen den Seitenketten, die gemeinsam das Protein in einer bestimmten Konformation halten. Viel stärker sind die Disulfidbrücken, kovalente Bindungen zwischen den Seitenketten.



Die Quartärstruktur
ist die Gesamtstruktur eines Proteins, die sich aus der Zusammenlagerung seiner Polypeptide (Untereinheiten) ergibt.

Abb. III) Übersicht: Die vier Ebenen der Proteinstruktur [Cam98]
Man kann die Strukturebenen in diesen schematischen Zeichnungen von Präalbumin unterscheiden, einem Blutprotein, das bestimmte Hormone und Vitamine transportiert.

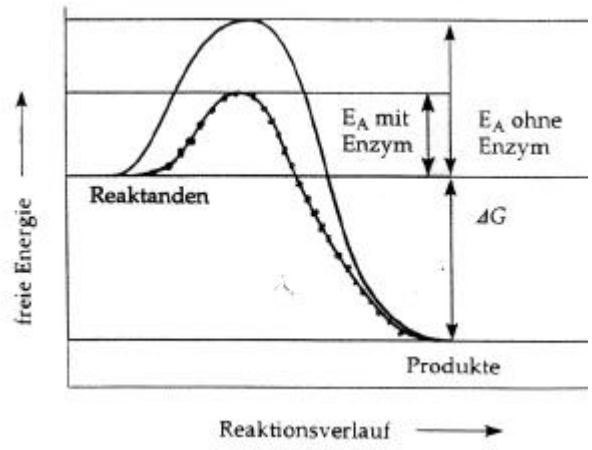
a) Die Primärstruktur ist die Reihenfolge der in einem Polypeptid kovalent verknüpften Aminosäuren.
b) Die Sekundärstruktur ist die Verformung eines Polypeptidrückgrats durch Ausbildung von Wasserstoffbrücken zu ∇-Helices und ®-Faltblättern.
c) Die Tertiärstruktur ist die Gesamtkonformation (Gestalt) eines Polypeptids, die durch Wechselwirkungen zwischen den Seitenketten der Aminosäuren stabilisiert wird.
d) Die Quartärstruktur ist die Assoziation von zwei oder mehr Polypeptiden (Untereinheiten) zu einem größeren Gebilde. Im Falle des Präalbumins besteht das gesamte Protein aus vier identischen Untereinheiten.



2.3 Enzyme – Ein Überblick

Enzyme sind katalytische Proteine. Ein Katalysator verändert die Reaktionsrate, ohne selbst in der Reaktion verbraucht zu werden. In Abwesenheit von Enzymen würde der chemische Verkehr durch die Stoffwechselwege in einem hoffnungslosen Stau enden.
Ein Enzym erhöht die Geschwindigkeit einer Reaktion, indem es die Aktivierungsenergie (E_A)- Barriere senkt, so dass der Übergangszustand schon bei gemäßigten Temperaturen erreicht wird (Abb. IV).
Ein Enzym kann aber die Änderung der freien Energie ($\cap G$) einer Reaktion nicht verändern; es kann keine endergonische Teilreaktion zu einer exergonischen machen. Enzyme können nur solche Reaktionen beschleunigen, die letztlich sowieso ablaufen würden, aber diese Funktion ermöglicht der Zelle einen dynamischen Stoffwechsel. Da Enzyme außerdem bei den von ihnen katalysierten Reaktionen sehr selektiv sind, bestimmen sie, welche chemischen Prozesse zu jedem Zeitpunkt in einer Zelle ablaufen.

Abb. IV) Enzyme: Herabsetzung der Aktivierungsbarriere [Cam98]
Die schwarze Kurve zeigt den Verlauf der Reaktion ohne Enzym, die rote (hier gestrichelt) den Verlauf mit Enzym.



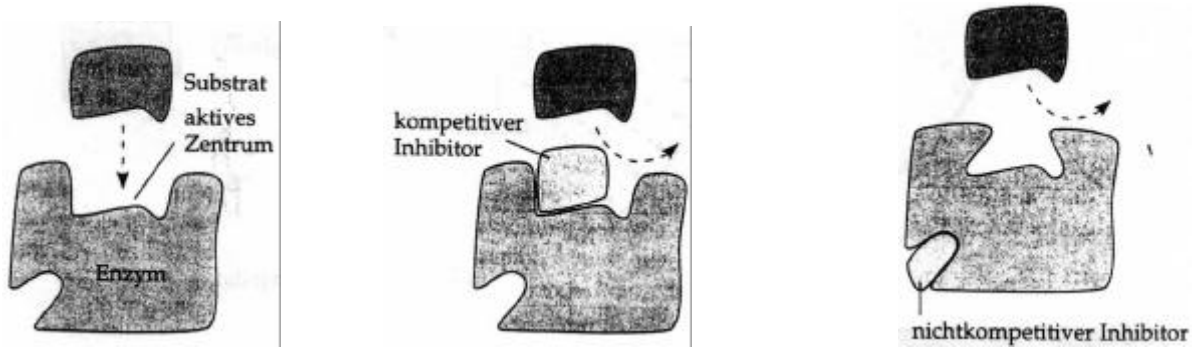
Der Reaktand, auf den ein Enzym einwirkt, wird als sein Substrat bezeichnet. Das Enzym bindet sein Substrat (bzw. Substrate). Am Enzym wird durch dessen katalytische Wirkung das Substrat in das Produkt (bzw. Produkte) umgewandelt. Enzyme sind:

→ substratspezifisch – nur spezifische Substrate werden gebunden
→ reaktionsspezifisch – jeder Enzymtyp katalysiert nur eine bestimmte Reaktion

Wodurch erfolgt diese molekulare Erkennung? Die Spezifität eines Enzyms beruht auf seiner Gestalt. Nur eine bestimmte Region des Enzymmoleküls bindet tatsächlich an das Substrat. Diese Region wird als das aktive Zentrum bezeichnet und ist typischerweise eine Tasche oder Spalte auf der Proteinoberfläche. Normalerweise wird das aktive Zentrum nur von einigen wenigen Aminosäure-Seitenketten des Enzyms gebildet, während der Rest des Proteinmoleküls das aktive Zentrum stabilisiert und schützt und die Regulation seiner Aktivität ermöglicht. Die Spezifität eines Enzyms beruht auf der Passgenauigkeit zwischen der Form des aktiven Zentrums und des Substrats.

Enzyminhibitoren:
Bestimmte Chemikalien hemmen selektiv die Wirkung spezieller Enzyme (Abb. V). Wenn der Inhibitor (Hemmstoff) sich kovalent (kovalente Bindung=Atombindung) an das Enzym bindet, ist die Hemmung normalerweise irreversibel. Bindet sich der Inhibitor dagegen über schwache Wechselwirkungen, ist die Inaktivierung reversibel. Viele Inhibitoren ähneln normalen Substratmolekülen und konkurrieren mit diesem um den Eintritt in das aktive Zentrum. Stoffe, welche die Aktivität von Enzymen herabsetzen, indem sie das Substrat an der Besetzung des aktiven Zentrums hindern, bezeichnet man als kompetitive Inhibitoren. In der Regel ist die *kompetitive Hemmung* reversibel und kann durch eine Erhöhung der Substratkonzentration überwunden werden, weil dann das Enzymmolekül statistisch viel häufiger mit einem Substrat- als mit einem Inhibitormolekül zusammentrifft.
Nichtkompetitive Inhibitoren beeinträchtigen enzymatische Reaktionen dadurch, dass sie sich an einen Teil des Enzyms binden, der weit vom aktiven Zentrum entfernt ist. Diese Wechselwirkung führt dazu, dass das Enzymmolekül seine Gestalt ändert, so dass das aktive Zentrum sein Substrat zwar noch bindet, es aber weniger effektiv in das Produkt umsetzt (*nichtkompetitive Hemmung*).
Beispiele:
Einige Enzyminhibitoren, die ein Organismus aus der Umwelt aufnimmt, wirken als Stoffwechselgifte. Die Pestizide DDT und Parathion zum Beispiel inhibieren Schlüsselenzyme des Nervensystems. Viele Antibiotika sind Inhibitoren für bestimmte Bakterienenzyme: Penicillin zum Beispiel blockiert das aktive Zentrum eines Enzyms, das viele Bakterien für den Aufbau ihrer Zellwand benutzen.
Beispiele für Enzyminhibitoren als Stoffwechselgifte mögen den Eindruck erwecken, dass Enzymhemmung im allgemeinen abnorm und schädlich ist. Tatsächlich aber ist selektive Hemmung und Aktivierung von Enzymen durch natürlich in der Zelle vorkommende Moleküle (sog. Liganden) ein ausgesprochen wichtiger Mechanismus der Stoffwechselkontrolle (z.B. Allosterische Regulation, Kooperativität). [Cam98]

Abb. V) Enzymhemmung [Cam98]



Für die physiologische Funktion ist die Regulierbarkeit der Enzymaktivität von außerordentlicher Bedeutung. Ist beispielsweise die dreidimensionale Struktur des aktiven Zentrums eines Proteins genau bekannt, das in ein bestimmtes Krankheitsgeschehen involviert ist, können gezielt Substanzen mit passendem Design entwickelt werden, die genau in das aktive Zentrum dieses Proteins passen, um es beispielsweise blockieren zu können. Dieses gezielte Entwickeln von Wirksubstanzen am Reißbrett wird als „molecular modelling“ (siehe Abschnitt 4) bezeichnet.
Wie kommt man zur exakten, dreidimensionalen Struktur eines Proteins?

3. Strukturanalyse

Proteine kann man auf viele Weisen analysieren.
z. B. Primärstruktur (= Proteinsequenzierung):

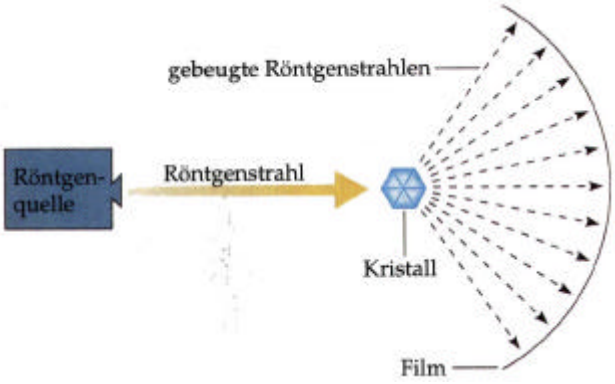
- Edman-Abbau
- Massenspektrometrie
- Ableitung aus cDNA-Sequenzen

Hier wird vorrangig auf die Methoden zur Bestimmung der exakten, dreidimensionalen Struktur eines Proteins eingegangen. Zwei Methoden werden angewendet: die Röntgenstrukturanalyse und die NMR-Spektroskopie (*nuclear magnetic resonance* = Kernmagnetresonanz).

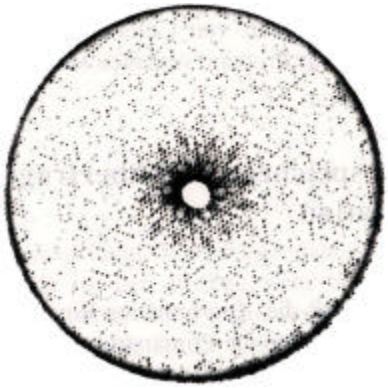
3.1 Röntgenstrukturanalyse

Die Methode zur Bestimmung der exakten, dreidimensionalen Struktur eines Proteins ist nach wie vor die Röntgenstrukturanalyse (Abb. VI). Röntgenstrahlen sind aufgrund ihrer sehr kleinen Wellenlänge (in diesem Fall von 0,154 nm) in der Lage, die Struktur von Proteinen bis zur atomaren Ebene aufzulösen. An den Atomen werden die Röntgenstrahlen gebeugt bzw. gestreut. Aus den Beugungsmustern kann auf die Elektronendichten und damit auf die Positionen der einzelnen Atome geschlossen werden. Voraussetzung für diese sehr potente Methode ist, dass die Proteine als Kristall vorliegen und damit die einzelnen Atome regelmäßig angeordnet sind. Vor allem membranintegrale Proteine und ganze Proteinkomplexe in eine Kristallstruktur „zu zwängen“ ist eine sehr diffizile und erst seit kurzer Zeit handhabbare Technik.

Abb. VI) Entwicklung eines Computermodells für die Struktur eines enzymatischen Proteins mit dem Namen Ribonuclease; vom Department of Biochemistry der Universität von Kalifornien, Riverside [Cam98]

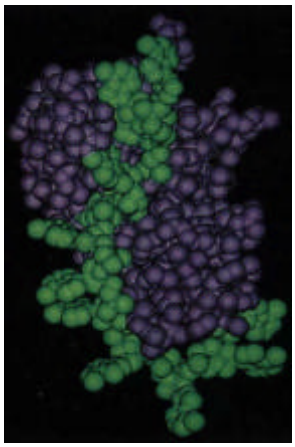


a) Röntgenkristallographie



b) Röntgenbeugungsmuster eines Proteinkristalls

c) Elektronendichtekarte



d) Computergraphikmodell des Proteins Ribonuclease (violett), das an einen kurzen Nucleinsäurestrang (grün) gebunden ist

3.2 NMR-Spektroskopie

Im Gegensatz zur Röntgenstrukturanalyse kann die NMR-Spektroskopie auch Strukturen von Proteinen auflösen, während sie sich in ihrem natürlichen Lösungsmittel Wasser befinden. Bei dieser Methode wird mit Hilfe extrem starker Magnetfelder – erzeugt von supraleitenden Magneten – die Rotation von Bestandteilen (Protonen) um den Atomkern beeinflusst. Nach einem Modell sind die Atomkerne elektrisch geladen. Die Protonen rotieren um den Atomkern (Spin). Bewegte Ladung baut ein Magnetfeld auf. Bei entsprechend starken externen Magnetfeldern können die Protonen parallel oder antiparallel (spin up, spin down → die zwei möglichen Richtungen) zu dem äußeren Magnetfeld rotieren. Da der antiparallele Spin energetisch aufwendiger ist als der parallele Spin, müssen die betreffenden Atomkerne die erforderliche Energie aufnehmen, indem sie elektromagnetische Strahlung im Bereich der Radiowellenfrequenz absorbieren.
Unter dem Einfluss eines äußeren elektromagnetischen Wechselfeldes gelangen die Protonen in einen angeregten Zustand, indem sie Energie aus dem Feld aufgenommen haben. Wenn sie aus dem angeregten Zustand zurückfallen, geben sie diese Energie wieder als Strahlung im Radiofrequenzbereich ab. Die Frequenz ist spezifisch für die Atomart und von der unmittelbaren Umgebung des Atoms abhängig. Somit kann die Molekülstruktur von Proteinen errechnet werden. Da die eigentliche Messzeit nur den Bruchteil einer Sekunde dauert, können auch Konformationsänderungen in Proteinen aufgelöst werden. Problematisch ist, dass die Kerne der wichtigsten biologischen Atome wie des Kohlenstoffs keine kernmagnetische Resonanz aufweisen. Sie müssen u. a. gentechnisch durch entsprechend kernmagnetresonanztaugliche Isotope des entsprechenden Elements ersetzt werden. Auch ist die Größe des zu untersuchenden Proteins stark begrenzt. Moderne NMR-Spektrometer, die mit einer Betriebsfrequenz von 750 oder 850 MHz arbeiten, können Proteinstrukturen bis maximal 30 kDa analysieren.

<u>Kleine Anregung:</u>
Aminosäuresequenz 3 + 3D-Struktur 3 = Ist das Problem der Struktur von Proteinen gelöst?!
Der rätselhafte Faltungscode der Proteine
Die Aufgabe, eine gewünschte dreidimensionale Tertiärstruktur eines Proteins mittels einer Aminosäuresequenz zielsicher zu konstruieren, wird zunehmend wichtiger. Ist z. B. die molekulare Ursache eines Defektes, der zu einem Krankheitsbild führt, eindeutig identifiziert, könnten mit Hilfe des gezielten Protein-Designs (*protein engineering*) passende molekulare „Schlüssel“ zum Beheben der molekularen Krankheitsursache gebildet werden. [Mu00]
Man könnte meinen, dass durch Korrelation der Primärstruktur vieler Proteine mit ihrer Konformation die Regel für die Proteinfaltung aufgedeckt werden könnten – insbesondere mit Hilfe von Computern. Leider ist das Proteinfaltungsproblem nicht so trivial. Die meisten Proteine durchlaufen auf ihrem Weg zu einer stabilen Konformation wahrscheinlich mehrere Zwischenstadien, und eine Betrachtung der „reifen“ Konformation sagt nichts über die Faltungsstadien aus, die für das Erreichen dieser Konformation notwendig sind. Mittlerweile haben Biochemiker allerdings Methoden zur Verfolgung von Proteinen durch die Zwischenstadien ihrer Faltung entwickelt. Außerdem haben Forscher die Chaperone entdeckt – Proteine, die andere Proteine bei ihrer Faltung unterstützen und dabei als zeitweilige Klammern wirken. [Cam98]

4. Molecular Modelling - Tertiärstruktur-Vorhersage von Proteinen

Warum Tertiärstruktur-Vorhersage von Proteinen?

- Hinweis auf Funktionsweise von Proteinen
- Leitstruktur für Mutagenese-Experimente
- Klassifizierung von Proteinen
- Vorlage für Wirkstoffdesign (nur sehr gute Modelle)
- Anhaltspunkt für Auswertung kristallographischer Rohdaten von Proteinen

4.1 Definition, Beispiel, Ziel

Def.:
Molecular Modelling ist eine Ansammlung von (computer-gestützten) Techniken für die Berechnung, Darstellung und Bearbeitung der realistischen dreidimensionalen Molekülstrukturen und ihren physikochemischen Eigenschaften. Als Rohmaterial dienen die „Rohdaten“ aus der Strukturanalyse. Das „Handwerkszeug“ stellen quanten- und molekülmechanische Rechenverfahren dar.[Sem02]

Die enorm verbesserten Methoden zur Aufklärung der Struktur und Funktion von Proteinen revolutionieren zur Zeit auch die Entwicklung neuer Medikamente.
Ein erfolgreiches Beispiel für molecular modelling
war und ist noch die Entwicklung neuer Medikamente zur Therapie der HIV-Infektion. Ziel dieser besonders potenten Medikamente ist eine HI-virale Protease, die die einzelnen Virusproteine aus einem langen Vorläufermolekül zurecht schneidet. Wenn dieses Enzym gehemmt wird, bleibt das Primärskript ungeschnitten und in der Folge können die einzelnen Virusproteine nicht mehr gebildet werden. Die Virusvermehrung wird unterbrochen. Es war klar, dass besonders solche Substanzen zur Hemmung der HIV-Protease in Frage kommen, die die Struktur einer Peptidbindung imitieren, schließlich ist die Aufgabe der Protease ja die Spaltung des viralen Vorläuferpolypeptids. Bald konnte gezeigt werden, dass das aktive Zentrum dieser Protease am effektivsten durch zyklische Harnstoffe besetzt wird. So gelang Wissenschaftlern innerhalb relativ kurzer Zeit die Entwicklung des ersten Proteaseinhibitors, der gezielt die HI-virale Protease blockiert. Die Proteaseinhibitoren erzielen heute in Kombination mit anderen anti-HI-viralen Substanzen sehr gute Erfolge bei der Reduzierung der Viruslast bei Infizierten. [Mu00]

Ziel des Molecular Modelling:

Ziel vom Molecular Modelling ist der Aufbau eines Struktur-Modells.
Hierzu nutzt man einerseits genaue Rechenverfahren (Quantenmechanik, Kraftfelder) und zusätzlich versucht man bei bekannten experimentellen Daten die Geometrie und physikochemische Eigenschaften zu extrapolieren. Man beginnt mit der Erzeugung eines Startmodells.
Um möglichst nah an den experimentellen Daten zu bleiben, sucht man in Datenbanken nach Molekülen mit sehr ähnlichen experimentellen Eigenschaften. Anschließend wird das Startmodell mithilfe von Kraftfeldmethoden oder Quantenmechanischen Berechnungen optimiert.[Sem02]

4.2 Methoden des Molecular Modelling

Die wichtigsten Ziele der Methoden des Molecular Modelling:
- Erstellung interaktiver Computergraphiken
- Modellierung kleiner Moleküle
- Molekülvergleiche
- Modellierung von Proteinen
- Modellierung von Protein-Ligand-Wechselwirkungen
- Ligandendesign

Für die effizienteste Lösung versucht man viel experimentelles Wissen (Molekül-, Kristall- und Proteinstrukturen) zu nutzen. In Datenbanken findet man findet man sowohl strukturelle als auch physikochemische Informationen (siehe Punkt 5. Proteindatenbanken).

4.2.2 Kraftfeldmethoden

Unter Kraftfeldmethoden versteht man empirische Verfahren zur Berechnung von Molekülgeometrien und Molekülenergien. Sie basieren auf der klassischen Newton'schen Mechanik und Elektrostatik. Man benutzt ein mechanisches Molekülmodell, in dem man die Atomen als Kugeln mit definierter Masse und die Bindungen (kovalente und nichtkovalente) als Federn betrachtet. Durch die Modellvorstellung wirken die Kraftfeldmethoden vereinfachend.
Ziel ist es die dreidimensionale Molekülstruktur mit minimaler Energie zu finden.

Die minimale Energie von Molekülen:
Die zentrale Idee ist, das Bindungslängen und Bindungswinkel versuchen Standardwerte einzunehmen (dort haben sie die minimale Energie). Durch sterische Wechselwirkungen (z.B. Abstoßung von benachbarten Gruppen) werden die Moleküle jedoch gezwungen teilweise in Bindungslängen oder Bindungswinkeln von ihren Idealwerten abzuweichen. Auf der Suche nach dem besten Kompromiss für das Molekül berechnet man die Energie verschiedener möglicher Konformationen, die sich aus folgenden Termen zusammensetzt:
Energie von ...
... Bindungsdehnung oder -stauchung (Bindungsstreckungen)
... Bindungswinkeldehnung oder -stauchung (Winkeldeformationen)
... Drehungen um eine Bindung (Veränderungen des Torsionswinkels)
... van-der-Waals-Wechselwirkungen (Anziehung und Abstoßung)
... elektrostatische Wechselwirkungen (Anziehung und Abstoßung)

Die einzelnen Terme werden unabhängig voneinander berechnet und anschließend zur Gesamtenergie aufsummiert. Anhand der Gesamtenergie kann man Aussagen über die Wahrscheinlichkeit der Existenz der jeweiligen Konformation machen.Wichtig für die richtige Berechnung sind hierbei außerdem die Wahl des richtigen Systems (Berücksichtigen der Lösungsmittel-Wechselwirkungen, in biologischen Systemen meist Wasser) und die Wahl der Startgeometrie (die Berechnung kann in einem lokalen Minimum "hängen bleiben" und findet das globale Minimum nicht).

Anwendbarkeit und Rechenaufwand:
Mithilfe der Molekülmechanik kann man fast alle organischen Verbindungen berechnen. Da der Rechenaufwand im Vergleich zu den quantenmechanischen Verfahren relativ gering ist, eignen sich Kraftfeldmethoden auch für Verfahren, die eine große Anzahl von Berechnungen erfordern:

- Berechnung von Molekülen in Solvensumgebung (Berücksichtigung der Wassermoleküle in der Umgebung)
- Konformationsanalyse (Berechnung mehrerer Konformationen eines Makromoleküls)
- Moleküldynamik
- Docking

Ein Nachteil der Kraftfeldmethoden ist, dass man von Annahmen ausgeht, die man nicht oder nur empirisch machen kann. Man erstellt sich Modelle. Die theoretisch korrekten Berechnungen sind die quantenmechanischen Berechnungen, da man hier keine Annahmen macht, aber die sind sehr aufwendig und können nur bis zu einer bestimmten Größe gemacht werden.

4.2.3 Quantenmechanische Berechnungen

Mithilfe der Quantenmechanik versucht man die elektronische Struktur von Molekülen zu beschreiben. Grundlage für diese Berechnungen ist die Schrödinger-Gleichung. Sie liefert aber nur Lösungen für sehr einfache Moleküle. Für Moleküle mit mehreren Elektronen gerät man in ein quantenmechanisches Vielteilchen-Problem (die Elektronen beeinflussen sich gegenseitig in ihrer Bewegung), das nur mit Hilfe von Näherungsverfahren gelöst werden kann. Am häufigsten wird dabei das Hartree-Fock-Verfahren verwendet.

Hartree-Fock-Verfahren:
Dieses Näherungsverfahren betrachtet die Bewegungen eines Elektrons im gemittelten Feld aller übrigen Elektronen und vernachlässigt damit die Korrelationen zwischen den Elektronen.
Damit wird das Vielteilchen-Problem auf ein Einteilchen-Problem zurückgeführt.
Der Zustand jedes Elektrons eines Moleküls wird durch eine Einteilchen-Funktion, das Atom- oder Molekülorbital beschrieben. Die Funktion für das ganze Molekül wird als antisymmetrisches Produkt der einzelnen Orbitale zusammengesetzt.

Man unterscheidet grundsätzlich 2 Verfahren bei quantenmechanischen Berechnungen: die ab-initio-Rechenverfahren und die semiempirischen Verfahren.

Ab-initio-Rechenverfahren und Semiempirische Methoden:
Bei den ab initio-Verfahren müssen keine Annahmen (wie z.B. der Hybridisierungsgrad bestimmter Atome) vor der Berechnung gemacht werden. Das ist günstig, weil man teilweise diese Annahmen aufgrund fehlender Kenntnisse über das Molekül gar nicht machen kann.
Es werden also keine empirischen Parameter in der Rechnung verwendet. Das macht diese Verfahren sehr genau. Sie haben aber dafür einen enormen Bedarf an Speicherplatz und Rechenzeit. Als Alternative dazu sind deshalb die semiempirischen Verfahren zu sehen. Hierbei werden einige der aufwendigen Schritte der ab initio-Verfahren durch einfache Näherungen ersetzt. Man spart dadurch ziemlich viel Rechenzeit.Insgesamt sind beide Verfahren zu rechenintensiv um sie auf große Moleküle anzuwenden.[Sem02]

Anwendungsgebiete Quantenmechanischer Verfahren

- Berechnung von Parametern für die Molekülmechanik
- Berechnung von Partialladungen und Elekronendichten und Molekülstrukturen
- Berechnung von Konformationsenergien kleiner Moleküle zum Kalibrieren von Kraftfeldern
- Berücksichtigung von Verschiebungen der Elektronendichte durch den Einfluss benachbarter Gruppen (induzierte Dipole)
- Beschreibung von chemischen Reaktionen

4.2.4 Homology Modelling, Threading, de novo-Strukturvorhersage

Homology Modelling
Da sehr viele Genom Sequenzen und 3D-Strukturen bereits aufgeklärt sind, steigt die Chance zu einer Aminosäuresequenz ein homologes Protein zu finden, dessen Struktur bekannt ist. Bei genügend Ähnlichkeit (> 70 % identische Aminosäuren), kann man versuchen, mittels Homology-Building für das unbekannte Protein ein Modell seiner Struktur zu berechnen.
z.B.: Programm SWISS-Model
→ http://www.expasy.ch/swissmod/SWISS-MODEL.html

Threading
Ausgehend von vorhandenen Protein-Folds versucht man eine Sequenz in alle möglichen Folds „einzupassen". Die Methode beruht auf evolutionär begrenzt entstandener Protein-Folds (ca. 1000). Somit steigt mit zunehmender Anzahl gelöster Strukturen die Erfolgschance, da sämtliche Protein-Folds bzw. Protein-Superfamilien (ca. 15.000) eines Tages in der Strukturdatenbank präsent sein werden.
→ http://www.hgmp.mrc.ac.uk/Registered/Option/threader.html (erfordert Registrierung)

***de novo*-Strukturvorhersage:**
Die Vorhersage beruht auf biophysikalischen Gesetzen der Proteinfaltung. Es ist nicht möglich alle theoretisch möglichen Strukturen zu berechnen. Um die wahrscheinlichsten Strukturen auszuwählen, benötigt man Algorithmen. Dies erfordert hoch parallele Rechner. Für kleine Peptide konnten bereits mit hohem Rechenaufwand Strukturen vorhergesagt werden. (siehe auch [Ko02, Fl02])

4.3 Datenbank – Konsequenzen

Aus den errechneten Daten aus der Strukturanalyse stellt sich nun die Frage nach den Anforderungen an eine Datenbank, die räumliche Datenstrukturen (3-dimensionale Punktdaten) speichert.
Die prinzipiellen Datenbankanforderungen sind Erhalt von Nachbarschaftsbeziehungen und möglichst Bildung von Clustern.

1. Durch welche Daten werden 3D-Strukturen von Proteinen beschrieben?
- Koordinaten (x,y,z) im 3D-Raum für alle Atome
- Atomart jedes Atoms
- Evt. Bindungen der Atome
- Evt. Wechselwirkungen der Atome
- Zusätzliche Informationen über das Protein

Eine Darstellung der 3D-Strukturen von Proteinen lässt sich also am besten durch 3-dimensionale Punktstrukturen (Punktdaten) beschreiben.

2. Durch welche Daten wird die Gestalt (Oberfläche) von Proteinen beschrieben?
- Zusammensetzung der Oberfläche aus Flächenstücken (Patches)
- die Flächenstücke selbst sind Ausschnitte aus Flächen 2. Ordnung (Eypsoide, Hyperboloide, Sattelstücke, o. ä.)
- Beschreibung der 2-dimensionalen Flächen durch Gleichungen (Punkte im n-dimensionalen Raum)

→ Speicherung der Parameter der Gleichungen in der Datenbank
- Datenbank:
  (1) Finde das Flächenstück, welches dort benannt wird
  (2) Rechne dort die entsprechenden Werte aus

Zunächst wird die Speicherung der errechneten räumlichen Datenstrukturen in Form von Punktdaten an einem einfachen Beispiel (Grid-File) betrachtet.

4.3.1 Speicherung räumlicher Datenstrukturen Punktdaten: Bucket-Methoden – Grid File (Gitter Datei)

Die Nutzung von Punktdaten ist bereits aus anderen Bereichen (z.B. Geographie, CAD) bekannt. Als Beispiel gelte ein zweidimensionaler Raum, um die Methode anschaulicher zu machen. Unsere Datenbank ist eine Sammlung von Aufzeichnungen (Daten), genannt Datei (*file*). Es gibt eine Aufzeichnung pro Datenpunkt. Die Datenbank enthält *N* Aufzeichnungen und *k* Schlüssel/Eigenschaften pro Stück.
Die einfachste Methode Punktdaten zu speichern ist in einer Liste, mit aufeinanderfolgen Daten. Beispiel soll eine Liste mit acht Städten und deren *x* und *y* Koordinaten sein (der Zusammenhang zwischen den Koordinaten und den Städten ist geographisch nicht korrekt, veranschaulicht die Methode aber besser).→ Tabelle 2

Tabelle 2 [Sa90]:

Name	x	y
Chicago	35	40
Mobile	50	10
Toronto	60	75
Buffalo	80	65
Denver	5	45
Omaha	25	35
Atlanta	85	15
Miami	90	5

Bei der Grid File Methode [Niev84, Hinr85a, Hinr85b]wird der Raum, in dem die Daten enthalten sind, in Regionen unterteilt, welche die Aufzeichnungen beinhalten. Diese Regionen bezeichnet man als *buckets*. Das Ziel dieser Methode ist mit höchstens zwei Plattenzugängen die Aufzeichnungen wiederzufinden und eine effiziente Handhabung der *range queries* (d. h. Regionssuche). Dies wird ermöglicht durch ein Gitter-Adreßbuch (*grid directory*), welches sich aus Gitterbausteinen (*grid blocks*) zusammensetzt. Alle Aufzeichnungen in einem Gitterbaustein sind in dem gleichen bucket gespeichert. Doch mehrere Gitterbausteine können sich ein bucket teilen, vorausgesetzt, die Gitterbausteine bilden ein k-dimensionales Rechteck in dem Raum mit den Aufzeichnungen.
Alle buckets bilden den Raum mit den Aufzeichnungen.
Der Zweck des Gitter-Adreßbuchs ist eine dynamische Korrespondenz zwischen den Gitterbausteinen in dem Raum mit den Aufzeichnungen und den bucket-Daten. Das Gitter-Adreßbuch setzt sich aus zwei Teilen zusammen. Der erste Teil ist ein dynamisches *k*-dimensionales Feld (*array*), welches einen Eingang für jeden Gitterbaustein enthält. Die Werte der Elemente sind Zeiger zu den relevanten bucket-Daten. Gewöhnliche buckets werden eine Kapazität von 10 bis 1000 Aufzeichnungen haben. Also ist der Eingang in das Gitter-Adreßbuch klein im Vergleich zu einem bucket. Der zweite Teil des Gitter-Adreßbuches ist ein Set von *k* ein-dimensionalen Feldern, genannt Linearskalen (*linear scales*). Die Linearskalen definieren eine Teilung der Bereiche jeder Eigenschaft. Sie ermöglichen den Zugang zu den passenden Gitterbausteinen durch Berechnung ihrer Adressen basierend auf dem Wert der relevanten Eigenschaften.
Jede Aufzeichnung wird wiedergefunden durch zwei Plattenzugänge – ein Plattenzugang je Gitterbaustein und bucket.

Abb. VII) Linearskalen für (a) x und (b) y im Zusammenhang zu Abb. VIII [Sa90]

x: 0 ξ 45 ξ 70 ξ 100
1 2 **3**
a

y: 0 ξ 42 ξ 70 ξ 100
1 **2** 3
b

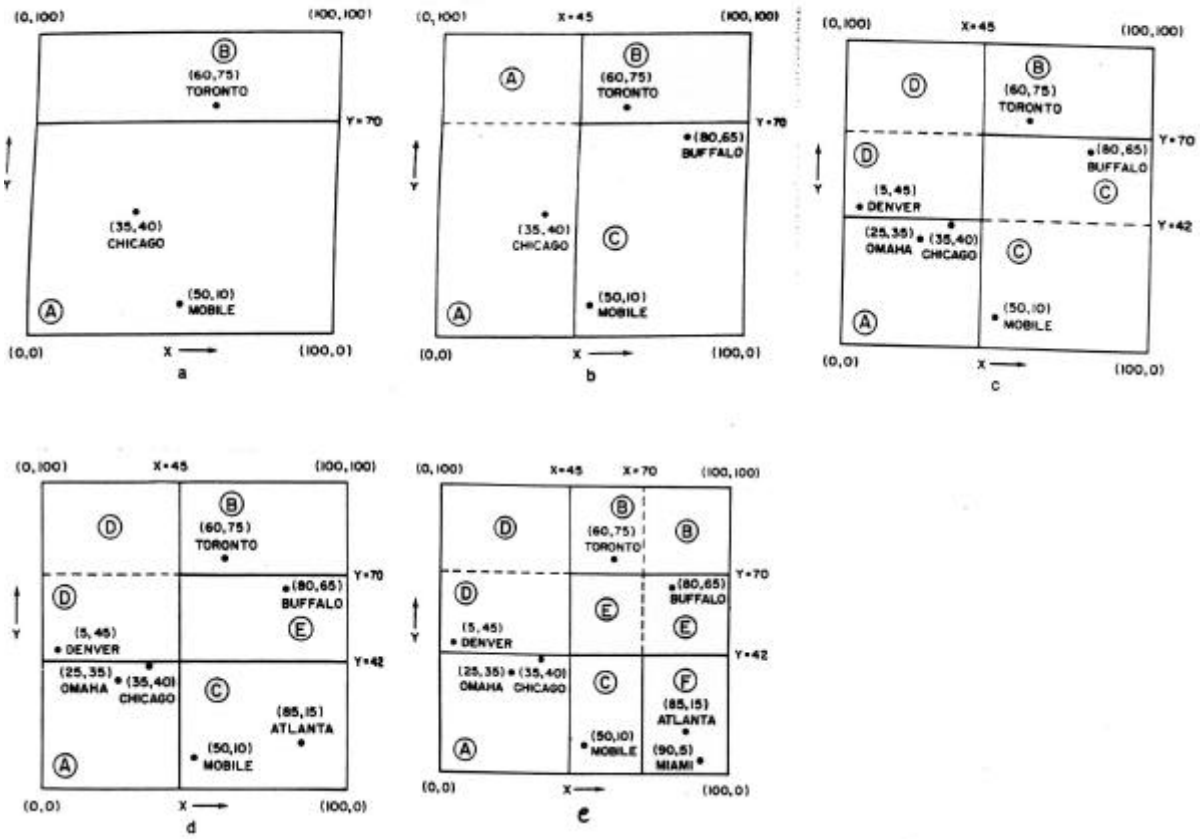
Beispiel: Punkt mit x=80 und y=65 (Ortskoordinaten von Buffalo)
→ der Punkt befindet sich im Gitterblock in Spalte 3 und Zeile 2 (siehe Abb. VIII)

Das Grid File ist teilweise attraktiv, weil es ein sehr schönes Wachstum zeigt, um so mehr Aufzeichnungen eingefügt werden. Laufen die buckets über, wird gesplittet. Das Resultat sind neue buckets und Bewegungen von Aufzeichnungen. Um den Splitting Prozess zu verstehen, hier ein Beispiel anhand der obenstehenden Tabelle (Tabelle 2):

<u>Bedingungen:</u>
- bucket-Kapazität: 2
- Es werden nach folgender Reihenfolge die Städte hinzugefügt:

1. Chicago
2. Mobile
3. Toronto
4. Buffalo
5. Denver
6. Omaha
7. Atlanta
8. Miami

Abb. VIII) Sequenz der Gitterteilung durch jeweiliges Hinzufügen von (a) Chicago, Mobile und Toronto, (b) Buffalo, (c) Denver und Omaha, (d) Atlanta, und (e) Miami [Sa90]



Das Einfügen von Chicago und Mobile führt dazu, daß der bucket A voll ist. Durch Hinzufügen von Toronto kommt es zu einem Überlauf von bucket A. Es wird gesplittet. Und zwar wird die *y*-Koordinate bei *y*=70 gesplittet, gleichzeitig wird die Linearskala modifiziert für die Eigenschaft *y*. Toronto ist nun in dem neuen bucket B (Abb. VIIIa). Als nächstes wird Buffalo hinzugefügt. Der bucket A ist voll, also wird die *x*-Koordinate bei *x*=45 gesplittet und die Linearskala in Bezug auf *x* modifiziert. Der neue Bucket C enthält nun Buffalo und Mobile. Mobile hat sich von bucket A in bucket C bewegt (Abb. VIIIb). Das Resultat dieses Splittings ist, dass beide Gitterbausteine (1, 1) und (1, 2) den bucket A teilen. Der Gitterbaustein (1, 2) ist lehr. Denver wird in bucket A eingefügt. Omaha ist ebenfalls in bucket A, so dass erneut gesplittet werden muß. Es wird die *y*-Koordinate bei *y*=42 gesplittet und die Linearskala in Bezug auf *y* modifiziert. Es entsteht der neue bucket D, welches nun Denver enthält. Da alle buckets konvex sein müssen, wird aus Gitterbaustein (1, 3) ein Teil von bucket D (Abb. VIIIc). Bucket C wird nun geteilt durch die Gitterbausteine (2, 1) und (2, 2). Nachdem Atlanta hinzugefügt wurde ist der bucket C voll. Der Gitterbaustein (2, 1) ist noch nicht voll. Dies führt zu einem neuen bucket E (Abb. VIIId). Es wurde jedoch keine erneute Teilung vorgenommen! Aber die Linearskala muss aktualisiert werden. Abschließend wird Miami in bucket C eingefügt, welcher voll ist. Also wieder Splitting der *x*-Koordinate bei *x*=70 und Modifizierung der Linearskala in Bezug auf *x*. Es entsteht der neue bucket F. Atlanta und Miami haben sich bewegt (Abb. VIIIe). Bucket B wird nun geteilt von den Gitterbausteinen (2, 3) und (3, 3), bucket E von den Gitterbausteinen (2, 2) und (3, 2).

Die Grid File Methode garantiert, dass jede Aufzeichnung mit zwei Plattenzugängen wiedergefunden werden kann. (siehe zu weiteren Methoden auch [Sa90])

4.3.2 Ähnlichkeitssuche

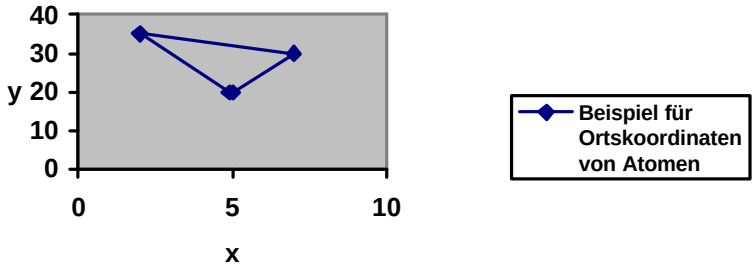
Die klassische Datenbankabfrage befasst sich mit der Suche nach Ähnlichkeit. Hier soll auf die Suche nach einer ähnlichen Form in einer Datenbank eingegangen werden.

1. **Berechnung des Schwerpunktes des Moleküls beispielhaft im zweidime nsionalen Raum**
   → dadurch werden mögliche verschiedene Orte der Atome eliminiert

$$x = \frac{\sum_{i=1}^{n} m_i \times x_i}{\sum_{i=1}^{n} m_i} \qquad y = \frac{\sum_{i=1}^{n} m_i \times y_i}{\sum_{i=1}^{n} m_i}$$

m_i...Masse der Atome
x_i, y_i...Orte der Atome
x, y...Schwerpunktkoordinaten
i...endlich viele Punkte

Abb. IX) Schwerpunktberechnung im zweidimensionalen Raum [Fa02]



Jeder Punkt (x_i, y_i) hat eine bestimmte Masse m_i. Nach oben beschriebener Formel können so z. B. hier für die drei Punkte die Schwerpunktkoordinaten x und y berechnet werden.

x	2	5	7
y	35	20	30

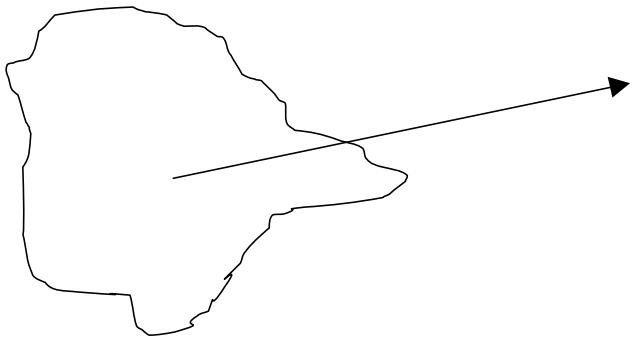
(2, 35)→ z.B. H (Wasserstoff) m_1=1,008u u...Atomare Masseeinheit
(5, 20)→ z.B. C (Kohlenstoff) m_2=12,01u u=1,66x10^{-27}kg
(7, 30)→ z.B. N (Stickstoff) m_3=14,007u

$$x = \frac{1{,}008\times 2 + 12{,}01\times 5 + 14{,}007\times 7}{1{,}008 + 12{,}01 + 14{,}007} = \underline{\mathbf{5{,}92}} \quad y = \frac{1{,}008\times 35 + 12{,}01\times 20 + 14{,}007\times 30}{1{,}008 + 12{,}01 + 14{,}007} = \underline{\mathbf{25{,}74}}$$

Ist der Schwerpunkt ermittelt, wird ein neues Schwerpunktkoordinatensystem erstellt. Den Ursprung dieses Koordinatensystems bildet der Schwerpunkt. Es handelt sich nun um ein Polarkoordinatensystem.

2. Zufällige Richtung vom Schwerpunkt aus festlegen

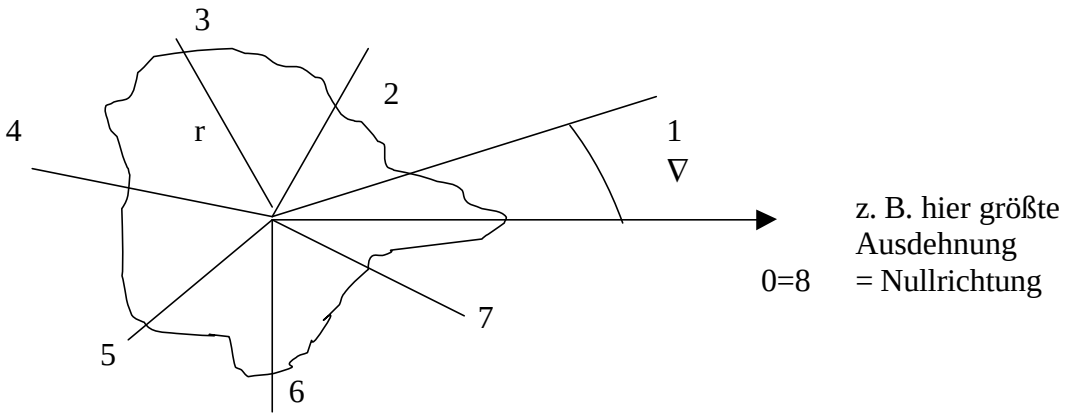
Abb. X) Beispiel für ein Molekül mit ermitteltem Schwerpunkt und Festlegung der Richtung:



3. Objekt „abtasten“

- größte Ausdehnung ist die Nullrichtung, dadurch werden Drehungen eliminiert
- Einteilung in Abschnitte (Winkel)→ Polarkoordinatensystem

Abb. XI) Beispiel:

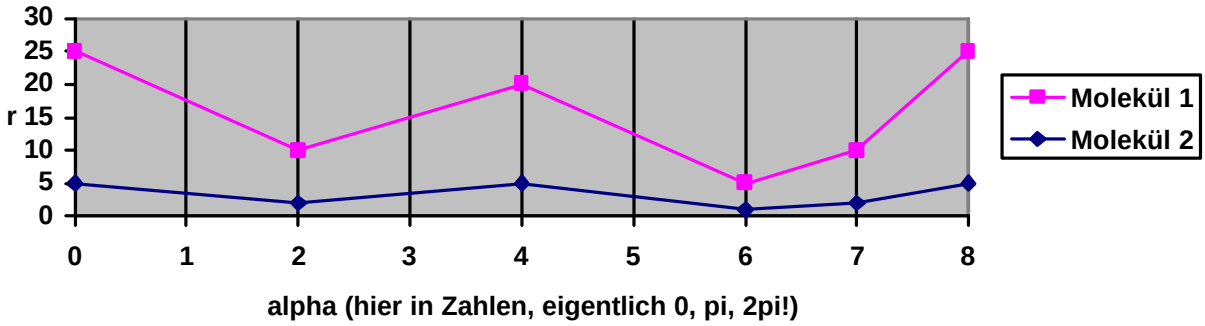


- Auftragung der Winkel im Koordinatensystem und Ermittlung der Schnittpunkte mit dem Molekül anhand der Orte der Atome (x_i, y_i)

4. Vergleich zweier Moleküle

- Auftragung der jeweiligen Werte r im Koordinatensystem

Abb. XII) Beispiel für den Vergleich zweier Moleküle



- Prüfen auf Proportionalität → proportional verändern, um proportionale Größenunterschiede zu eliminieren
- statistische Auswertung der Kurven, z. B. Regression (Berechnung der Quadratsummen) und Bedingungen für die Ähnlichkeit festlegen

5. Proteindatenbanken

Datenbanken dienen dem Beschreiben, Speichern und Wiedergewinnen von großen Datenmengen.
Eine Auswahl an verschiedenen Proteindatenbanken:

5.1 Sequenzdatenbanken (=primäre Datenbanken) [Fl02]

SWISS-PROT
- unterhalten vom EBI und SIB (Swiss Institute for Bioinformatics)
- Annotationen durch Kuratoren, d. h. viele und gut abgesicherte Informationen zum Protein (Annotation = Beschreibung von Merkmalen des Proteins)
- geringe Redundanz
- relativ geringe Anzahl von Einträgen: 116.776 am 25.10.02
- Datenquellen:
    - direkte Einreichungen
    - Übersetzung von CDS aus EMBL
    - Übernahme aus PIR

TrEMBL
- Computer-annotiertes Supplement von SWISS-PROT
- Übersetzungen aller CDS in EMBL (TrEMBL = translated EMBL)
- Sektion SP-TrEMBL: für Aufnahme in SwissProt vorgesehen, aber noch nicht durch Kuratoren annotiert
- Einträge: 680.075 Einträge am 25.10.02
- Sektion REM-TrEMBL (remaining TrEMBL): Sequenzen, die nicht in SwissProt aufgenommen werden sollen, z.B. Ig's, MHCs, kurze oder hypothetische Fragmente
- automatische Übersetzung der amerikanischen Datenbank GenBank heißt GenPept

PIR = Protein Information Resource
- unterhalten von PIR-International, einem Konsortium aus einem amerikanischen, einem japanischen und einem europäischen Institut
- europäischer Partner: MIPS (Munich Information Center for Protein Sequences)
- Basiert auf dem *Atlas of Protein Sequence and Structure* von Margaret Dayhoff
- Einteilung der Proteine in Superfamilien, Familien etc.
- Annotation durch Kuratoren/Experten und anhand dessen auch automatisch (d.h. Experten-Annotationen werden auf Proteine derselben Familie übertragen)
- Einträge: 283.227 am 21.10.02 (283.174 am 02.04.02)

→ Zugang zu SWISS-PROT, TrEMBL und PIR z.B. über SRS beim DKFZ

Zusammengesetzte Datenbanken
- Problem: einzelne Datenbanken könnten unvollständig sein; um nicht mehrere Datenbanken durchsuchen zu müssen, gibt es zusammengesetzte Datenbanken
- NRDB = Non-Redundant DataBase; enthält GenPept, SWISS-PROT, PIR u.a.; unterhalten vom NCBI, Standarddatenbank für Entrez Protein-Anfragen
- PIR-NREF = PIR Non-redundant Reference Database; enthält PIR, SWISS-PROT, TrEMBL u.a.; unterhalten von PIR-International
- SPTR (auch: SWall; auch: SP-TrEMBL); enthält SWISS-PROT und TrEMBL

- In allen Fällen werden doppelte Sequenzen eliminiert
- ABER: zusammengesetzte Datenbanken sind dennoch redundant, da z. B. sehr ähnliche, aber nicht identische Sequenzen (Polymorphismen, Sequenzierfehler) nicht eliminiert werden

UniProt – United Protein Database
- Lt. EBI-Pressemitteilung vom 23.10.02 werden SWISS-PROT, TrEMBL und PIR zu einer gemeinsamen Datenbank zusammengefügt
- Wichtiges Ziel:
  Verbesserung der automatischen Annotation durch Kombination von hauseigenen Verfahren

5.2 Motivdatenbanken (=sekundäre Datenbanken) [Fl02]

- entstehen durch Analyse primärer Datenbanken
- enthalten Sequenzmotive, die häufig mit einer biologischen Funktion assoziiert sind
- Ziel: durch Vergleich einer Sequenz mit der Motiv-Datenbank Rückschlüsse auf die Funktion des unbekannten Proteins ziehen (sehr wichtig auch für automatische Annotationen)
- Beispiele für sekundäre Datenbanken:

PROSITE @ http://www.expasy.ch/prosite/
PRINTS @ http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
BLOCKS @ http://www.blocks.fhcrc.org/
Pfam @ http://www.sanger.ac.uk/Software/Pfam/
PRODOM @ http://prodes.toulouse.inra.fr/prodom/doc/prodom.html

PROSITE
- Unterhalten vom Swiss Institute of Bioinformatics (SIB)
- Idee: Charakterisierung von Proteinfamilien anhand der konserviertesten Sequenzabschnitte
- enthält Motive in Form von patterns (Muster) und profiles (Profile)
- patterns beschreiben kurze Motive (10 - 20 Aminosäuren)
- sehr kurze Motive werden als Sites bezeichnet
- profiles beschreiben ganze Sequenzen als Motive
- Erstellung von PROSITE-patterns:
  - Multiple Alignment mit bekanntermaßen verwandten Sequenzen aus Swiss-Prot
  - Auswahl konservierter Bereiche und Ableitung einer Konsensus-Sequenz (seed-pattern)
  - Durchsuchen von Swiss-Prot mit dem seed-pattern
  - Prüfen des Ergebnisses auf falsch-positive und falsch-negative Treffer
  - Anpassung des patterns und erneute Datenbanksuche
  - Wiederholung bis optimales pattern gefunden wurde
- Schreibweise von PROSITE-patterns:
  - Aminosäuren im Ein-Buchstaben-Code
  - einzelne Positionen sind durch Striche - getrennt
  - eckige Klammern, z.B. [AS]: A oder S an dieser Position
  - x: irgendeine Aminosäure an dieser Position
  - geschweifte Klammern, z.B. {PG}: alle Aminosäuren sind möglich, aber nicht P und G

- Zahlenangaben hinter Buchstaben oder Klammern, z.B. [FY]4: Angabe bezieht sich auf die angegebene Anzahl aufeinanderfolgender Positionen; Angabe von Zahlen in Klammern, z.B. x(2,4): 2 bis 4 beliebige Aminosäuren

- Beispiele für patterns:
  - N-Glykosylierungsstelle:

    N-{P}-[ST]-{P} [N is the glycosylation site]

  - Zink-Finger (in DNA-bindenden Proteinen):

    C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H [The two C's and two H's are zinc ligands]

- ACHTUNG: Übereinstimmung mit einem Motiv bedeutet nicht zwangsläufig, dass es in diesem Protein auch funktionell ist! Immer die in PROSITE ausführliche Dokumentation beachten!
- Probleme mit patterns:
  - Es werden nur exakt mit dem pattern übereinstimmende Sequenzen gefunden
  - Aufweichung der patterns führt, besonders bei kleinen patterns, zu sehr vielen Treffern

  $\Rightarrow$ patterns eignen sich nicht zur Erkennung entfernt verwandter Proteine
- profiles in PROSITE
  - Profile beschreiben über die gesamte Länge eines multiplen Alignments die Wahrscheinlichkeit, dass an einer Position eine bestimmte Aminosäure, eine Insertion oder eine Deletion auftritt

in PROSITE gibt es deutlich mehr patterns als profiles

PRINTS

- enthält Fingerprints; das sind mehrere Motive, die eine Proteinfamilie charakterisieren; Einträge haben die Form lokaler multipler Alignments

BLOCKS

- Einträge enthalten ebenfalls mehrere lokale multiple Alignments zur Charakterisierung von Proteinfamilien; automatisch berechnet ausgehend von Proteinfamilien in PROSITE; Grundlage für die BLOSUM-Substitutionsmatrizen

Motiv-Datenbanken,die auf HMMs beruhen

- HMM = Hidden Markov Modell; statistisches Modell zur Beschreibung von Übergangswahrscheinlichkeiten
- ein HMM eines multiplen Alignments beschreibt für jede Position die Wahrscheinlichkeit, dass eine bestimmte Aminosäure, eine Deletion oder eine Insertion auf die vorhergehende Position folgt
- Datenbank mit HMMs für multiple Alignments: Pfam

ProDom

- Datenbank von Proteindomänen
- Domänen = Proteinabschnitte, die während der Evolution in verschiedenen Proteinen verwendet wurden

InterPro

- Datenbank, die Motive u.a. aus PROSITE, PRINTS und PFAM sowie Domänen aus ProDom enthält
- erlaubt simultane Suche in allen gängigen Motivdatenbanken

  @ http://www.ebi.ac.uk/interpro/

5.3 weitere Datenbanken (z. B. Stoffwechsel, genetische Karten, Erbkrankheiten/ Mutationen, Transkriptionsfaktoren und ihre Bindungsstellen)

- Stoffwechsel: KEGG = Kyoto Encyclopedia of Genes and Genomes (unterhalten von GenomeNet)
- genetische Karten: LocusLink (unterhalten vom NCBI)
  Erbkrankheiten/Mutationen: OMIM (Online Mendelian Inheritance in Man)
- Transkriptionsfaktoren und ihre Bindungsstellen: Transfac (z.T. frei zugänglich, z.T. nur mit Lizenz ⇒ Firma Biobase in Braunschweig)

5.4 Strukturdatenbanken [Fl02]

- **PDB** = Protein Data Bank

→ Zentrale Datenbank für Proteinstrukturen
→ unterhalten vom Research Collaboratory for Structural Bioinformatics (RCSB)
→ enthält auch DNA- und Zuckerstrukturen

@ http://www.rcsb.org/pdb/

→ PDB-Dateien:
- Enthalten nur Atom-Koordinaten
- Bindungen werden vom Viewer-Programm anhand von Regeln (chemistry rules) berechnet;
  z.B. C-C-Bindung = 1,5 Angström

→ Viewer für Strukturdateien
- RasMol
- Chime (schnelles RasMol-Derivat)
- Protein Explorer (Weiterentwicklung von Chime); schnell und einfach zu bedienen; funktioniert auch online, wenn man das Chime-Plugin hat

@ für alle 3 Programme (und das Chime-Plugin):
http://www.umass.edu/microbio/rasmol/

@ oder Links aus PDB-Einträgen
- Swiss-PDB Viewer (Deep View); professionell, sehr viele Darstellungsmöglichkeiten, Dateien können für PovRay exportiert werden

@ http://www.expasy.ch/spdbv/

@ sehr gutes Tutorial: http://www.usm.maine.edu/~rhodes/SPVTut/index.html

- **MMDB** = Molecular Modelling Database

→ unterhalten vom NCBI (Synonym: NCBI Structure Division)
→ enthält die Daten aus der PDB, aber in einem anderen Format und mit zusätzlichen Informationen
→ Zugang über NCBI → Entrez → Structure

@ http://www.ncbi.nlm.nih.gov/Structure/

→ MMDB-Dateien:
- Werden aus PDB-Dateien erstellt
- Prinzip: Proteinsequenz definiert schon die chemische Struktur der Polypeptidkette
- Bindungen werden vom Viewer-Programm aus ‚residue dictionaries' abgerufen
- Dateien enthalten auch Informationen über Struktur- und Sequenz-Nachbarn

→ Viewer für Strukturdateien
- Cn3D

@ Link zum Download erscheint in Einträgen

- **Datenbanken zur strukturellen Klassifizierung z.B. SCOP**

→ = Structural Classification of Proteins
→ Konzept: Protein-Domänen sind die Basiseinheiten der Proteinevolution
→ Eine Domäne ist ein Teil des Proteins, der sich (mehr oder weniger) unabhängig zu einer (mehr oder weniger) kompakten Struktur faltet
→ Zur Klassifizierung werden Proteine in Domänen zerlegt
→ Hierarchische Klassifizierung in SCOP:
- Familie: homologe Proteine/Domänen mit ähnlicher Sequenz, Struktur, evt. Funktion
- Superfamilie: ähnliche Struktur, evt. Funktion, Sequenzähnlichkeit kann gering sein
- Fold: gleiche Anordnung von Sekundärstrukturelementen (es gibt nur eine begrenzte Anzahl natürlich vorkommender Faltungsmotive oder Folds, derzeit gibt es ca. 500 bekannte Folds)
- Klasse: generelles Vorkommen von Sekundärstrukturelementen

→ Klassen:
- α: all alpha = nur α-Helices
- β: all beta = nur β-Faltblätter
- α/β: alpha and beta = einzelnes, paralleles Faltblatt, zwischen den Strängen α-Helices (β-α-β)
- α+β: alpha and beta = antiparallele Faltblätter, Stränge durch Hairpins verbunden, α- und β-Anteile getrennt

@ http://scop.mrc-lmb.cam.ac.uk/scop/

Zusammenfassung

Die Biologen versuchen über die Aufklärung der Proteine Stoffwechselwege zu beeinflussen. Um die Daten über die Proteinanalysen auszuwerten benötigen sie eine effektive Speicherung der Daten und eine Vergleichbarkeit (Ähnlichkeitssuche) der Daten. Was die Biologen genau wollen, ist nicht ganz eindeutig, was eine effektive computer-gestützte Auswertung schwer macht. Das erstellen einer Datenbank mittels Datenbank Management Systemen (DMBS) ist ein langwieriger Prozess und unterliegt somit auch dem rasanten Fortschritt.
Deshalb beruhen heute noch viele Bio-Datenbanken aus indizierten ASCII Textdateien, den sog. „flat files".
In der Ausarbeitung sind zwei möglich Ansätze (Speicherung als Punktdaten, Ähnlichkeitssuche) für die Anwendung in DMBS gezeigt, um dem Problem der Klärung der 3D-Struktur aus Sicht der Informatik näher zu kommen und vielleicht dazu beitragen eine effektive Lösung für die Anlayse von 3D-Strukturen von Proteinen zu finden.

6. Literaturverzeichnis

[Cam98] Campbell, Neil A.: Biologie. Spektrum Verlag, 1998

[Fl02] http://www.fh-flensburg.de/studierende/deutsch_4015.html

[Ko02] http://www.sfb363.uni-halle.de/kurs/3D.html

[Mu00] Munk, Katharina: Grundstudium Biologie/Biochemie, Zellbiologie, Ökologie, Evolution. Spektrum Verlag, 2000

[Niev84, Hinr85a, Hinr85b] siehe Literaturverzeichnis von [Sa90]

[Sa90] Samet, Hanan: The Design and Analysis of Spatial Data Structures. Addison-Wesley, 1990

[Sem02] http://www.Seminar_Biodatenbanken\BIOINFORMATICS MEETS CHEMISTRY - Molecular Modelling2.htm

Abbildungs- und Tabellenverzeichnis

Tabellen:

Abbildungen:

Glossar

Dalton (Da):	Die Atommasseeinheit; ein Maß für die Masse von Atomen (oder Molekülen) und subatomaren Einheiten. (kDa=kiloDalton)
Dipol-Dipol-Kräfte:	Anziehungskräfte zwischen polaren Molekülen aufgrund der Anziehung zwischen entgegengesetzten Polen
Fold:	gleiche Anordnung von Sekundärstrukturelementen
kalibrieren:	auf genaues Maß bringen
Konformation:	Gestalt
kovalente Bindung:	Atombindung
Kristallgitter:	Dreidimensionales, regelmäßiges Muster von Punkten, die Lagen gleicher Umgebung repräsentieren.
Ligand:	Ein Molekül, das spezifisch an einen Rezeptorberiech einen andern Moleküls bindet.
Rezeptoren:	In der Plasmamembran oder im Cytoplasma befindliche oder auch frei bewegliche Moleküle (meist Proteine), die durch Bindung von Molekülen eine Reaktion auslösen.
Polypeptidkette:	Polymere aus Aminosäuren
Solvens:	Lösungsmittel
Substrat:	Reaktand, auf den ein Enzym einwirkt