Biomolekulare Informatik als Anwendungsbereich für Hochleistungsrechner

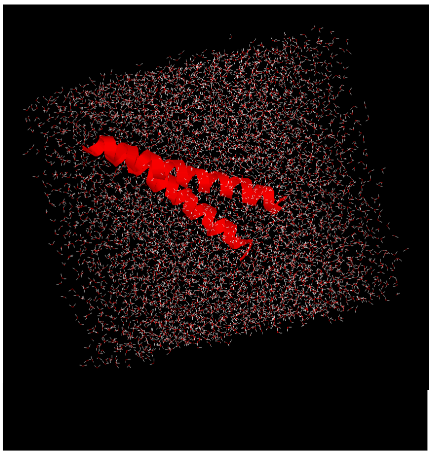
Heinrich Sticht, Gabi Seidel, Dieter Willbold & Paul Rösch
Lehrstuhl für Struktur und Chemie der Biopolymere, Universität Bayreuth, D-95448 Bayreuth

Freie Moleküldynamik

Moleküldynamiksimulationen ohne experimentelle Randbedingungen bieten die Möglichkeit, die Stabilität von dreidimensionalen Strukturen zu überprüfen. Darüber hinaus ermöglicht diese Methode Aussagen über die Zeitskala und Amplitude von Bewegungen, die in Biomolekülen auftreten. Für zuverlässige Simulationen ist eine explizite Berücksichtigung des Lösungsmittels unbedingt notwendig, was zu einer durchschnittlichen Größe der Systeme von 20000-30000 Atomen führt und diese Methode rechnerisch sehr aufwendig macht.
Die Simulationsdauer richtet sich nach der Art der zu charakterisierenden Bewegungen: So können Bewegungen von Aminosäuren-Seitenketten oder flexiblen Loops bereits auf einer Pikosekunden- bis Nanosekunden-Zeitskala beobachtet werden. Für biologisch wichtige Effekte wie Bewegungen von Proteindomänen oder Wechselwirkungen mit Liganden sind dagegen Simulationszeiten von mehreren hundert Nanosekunden erforderlich.

Projekte:
- Untersuchung der Proteinstruktur in Abhängigkeit vom Lösungsmittel am Beispiel des retroviralen Transaktivators aus dem Pferdeanämievierus (EIAV-Tat) (Sticht et al., 1994)
- Untersuchung der konformationellen Stabilität des Leucin-Zippers aus DIP (Seidel et al., 1997)
- Untersuchung der Thermostabilität von Proteinen am Beispiel von [4Fe4S]-Ferredoxinen (Sticht et al., 1996)

Notwendiger Rechnerbedarf für eine Simulation von einer Nanosekunde:
- cpu-Zeit: 126 Tage (HP9000/735-Workstation, 99 MHz cpu)
- Hauptspeicher: > 64 Mbyte
- Plattenspeicher: ca. 4 Gbyte



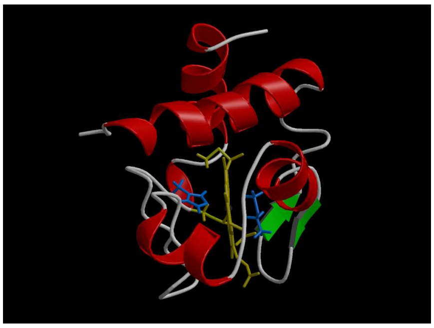
Ergebnis einer freien Moleküldynamik-Simulation des Leucin Zippers aus DIP. Die Helices des dimeren Proteins sind als rote Bänder dargestellt. In dieser Simulation wurden 8000 Wassermoleküle berücksichtigt (als Stäbchen eingezeichnet).

Strukturberechnung aus NMR-Daten

Die Auswertung der verschiedenen NMR-Experimente liefert eine Vielzahl an Informationen über die Konformation des Proteins. Um die erhaltenen Daten in die dreidimensionale Struktur umsetzen zu können, werden Distanzgeometrie und Moleküldynamik als Methoden eingesetzt.
Von der nativen Struktur des Proteins wird angenommen, daß sie die energieärmste Konformation darstellt. Das Auffinden eines derartigen globalen Energieminimums ist jedoch erheblich durch die Existenz vieler lokaler Energieminima erschwert. Ein generelles Problem jeder Strukturberechnung ist daher, das globale Minimum der Potentialfunktion, die von sehr vielen Parametern abhängt, zu finden. Eine Lösung für ein derartig komplexes, nicht-lineares Optimierungsproblem stellt die Methode des "simulated annealing" dar.
Um zu überprüfen, ob man mit einer Berechnung tatsächlich eine Konformation erzeugt hat, die nahe am globalen Minimum liegt, muß jede Simulation für ca. 50-100 verschiedene Startstrukturen durchgeführt werden. Da die Konvergenz aller im Moment verfügbaren Algorithmen noch unzureichend ist, müssen alle Strukturberechnungen darüber hinaus in 10-20 iterativen Schritten durchgeführt werden, wobei die Gewichtung der experimentellen Daten systematisch verändert wird.
Der hohe rechnerische Aufwand dieser Methoden kommt aber dadurch zustande, daß die vorhandenen Algorithmen für jede Fragestellung erneut auf ihre Brauchbarkeit hin überprüft und gegebenenfalls modifiziert werden müssen.

Projekte:
- Strukturbestimmung der retroviralen Transaktivatoren HIV-Tat und EIAV-Tat
- Strukturbestimmung des Cytochroms c6 und des [2Fe2S]-Ferredoxins aus *Synechococcus elongatus*
- Strukturbestimmung des Birkenpollenallergens Bet v1
- Strukturbestimmung des menschlichen Parathyroidhormons
- Strukturbestimmung der viralen Protein vpu, vpr und der nef-Funktionsdomäne

Notwendiger Rechnerbedarf einer Strukturberechnung aus NMR-Daten:
- cpu-Zeit: ca. 140 Stunden für jeden der 10-20 iterativen Schritte der Berechnung (Sun SparcUltra Workstation)
- Hauptspeicher: > 32 MByte
- Plattenspeicher: ca. 100 Mbyte



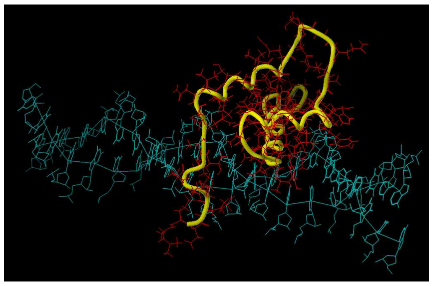
Schematische Darstellung der Struktur des Cytochroms c6 aus *Synechococcus elongatus*. Vier α-Helices (rot) und ein antiparalleles β-Faltblatt (grün) sind als Hauptelemente der Sekundärstruktur eingezeichnet. Das Häm (gelb) und die beiden axialen Hämliganden (blau) sind als Stäbchen dargestellt.

Komparatives Modellieren

Einige Biomoleküle sind einer Strukturbestimmung mit Röntgenkristallographie oder NMR-Spektroskopie nicht zugänglich, da sie weder kristallisieren, noch in millimolaren Konzentrationen löslich sind. Für solche Moleküle kann das komparative Modellieren eine Alternative zu einer Strukturbestimmung bieten.
Komparatives Modellieren ist eine Methode, die auf der Basis von bekannten dreidimensionalen Proteinstrukturen in der Lage ist, die Struktur von unbekannten Proteinen mit ähnlicher Sequenz vorherzusagen. Die Anwendung dieser Methode ist möglich, da die Strukturbestimmung von über 1000 Proteinen gezeigt hat, daß Proteine mit ähnlicher Sequenz meist auch eine ähnliche dreidimensionale Struktur besitzen. Bei einer Sequenzidentität von mehr als 40% zu einer bekannten Struktur kann diese Methode Strukturmodelle liefern, deren Qualität der einer niedrig aufgelösten Kristallstruktur gleichkommt. Energieminimierung und Moleküldynamik werden häufig als Methoden eingesetzt, um die Plausibilität der erhaltenen Strukturmodelle zu überprüfen und die erhaltenen Strukturen zu verfeinern. Im Unterschied zur 'freien' Moleküldynamik ergibt sich hier ein hoher Rechenbedarf nicht nur durch die Länge der Simulationszeit, sondern auch dadurch, daß für jede Modellierung mehrere Strukturmodelle berechnet und analysiert werden müssen. Um ein zuverlässiges Strukturmodell zu erhalten, müssen dabei je nach Fragestellung ca. 100 bis 10000 Strukturen erzeugt und anschließend auf ihre Übereinstimmung mit den verfügbaren experimentellen und empirischen Daten überprüft werden.

Projekte:
- Erstellen eines Strukturmodells der Adenylosuccinat-Synthetase aus *S. cerevisiae* (Sticht et al., 1997)
- Modellierung von Eisen-Schwefel-Clustern (Sticht et al., 1996; Baumann et al., 1996) und Hämgruppen (Beißinger et al., 1997) aus Metalloproteinen.
- Entwicklung von Methoden zur Vorhersage und Verfeinerung der Struktur von modulären Proteinen (Strukturuntersuchungen der Typ II Module aus Fibronectin (Sticht et al., 1997) und eines CP-Modulpaars des sp35 Proteins aus *vaccinia virus*)
- Modellierung der Protein-Nukleinsäure Komplexe des HIV-Tat und EIAV-Tat

Notwendiger Rechnerbedarf für das Erzeugen und Analysieren einer Schar von 1000 Strukturen:
- cpu-Zeit: 1500 Stunden (Sun SparcUltra Workstation)
- Hauptspeicher: > 64 MByte
- Plattenspeicher: ca. 500 Mbyte



Strukturmodell eines EIAV-Tat-DNA-Komplexes. Das Transaktivatorprotein des Pferdeamämievirus (EIAV) ist in rot und die DNA in blau dargestellt. Als Grundlage für das Modellieren dieses Komplexes diente die Struktur des engrailed-Homöodomänen-DNA Komplexes.

Datenbanken

Für die meisten strukturellen Daten im Bereich der Biochemie existieren heute Datenbanken. Sie beinhalten Informationen über Gensequenzen, Aminosäuresequenzen oder dreidimensionale Strukturen von Biomolekülen. Datenbanken werden z.B. eingesetzt, um Informationen über die Homologie zwischen verschiedenen Proteinen zu erhalten, oder auch, um bei bekannter Gen- oder Aminosäuresequenz die Funktion des Proteins vorherzusagen. Datenbank-Analysen stellen meistens einen Schritt dar, der der Strukturbestimmung vorhergeht. Der notwendige cpu-Bedarf richtet sich dabei nach der Fragestellung und der Größe der entsprechenden Datenbank. Für ausführlichere Suchen sind einige Stunden cpu-Zeit notwendig. Das gegenwärtig exponentielle Wachstum der Datenbanken läßt jedoch erwarten, daß auch der Rechnerbedarf in nächster Zeit stark ansteigt.

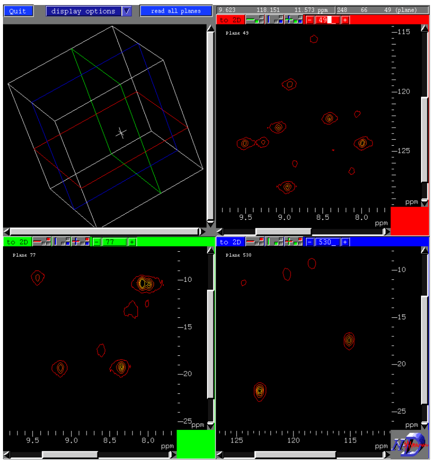
```
PRIO_BOVIN  MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYPPQGGGGW
PRIO_HUMAN  --MANLGCWMLVLFVATWSDLGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYPPQGGG--
PRIO_MESAU  --MANLSYWLLALFVAMWTDVGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYPPQGGG--
PRIO_MOUSE  --MANLGYWLLALFVTMWTDVGLCKKRPKPGG-WNTGGSRYPGQGSPGGNRYPPQGG---
PRIO_RAT    ------------------------------GG-WNTGGSRYPGQGSPGGNRYPPQSGG--
PRIO_SHEEP  MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYPPQGGG--
BETA_AMYLO  ------------------------------------------------------------

PRIO_BOVIN  GQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGGWGQGGTHGQWNKPSKPKTNM
PRIO_HUMAN  ------GWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNM
PRIO_MESAU  ------TWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHNQWNKPSKPKTNM
PRIO_MOUSE  ------TWGQPHGGGWGQPHGGSWGQPHGGSWGQPHGGGWGQGGGTHNQWNKPSKPKTNL
PRIO_RAT    ------TWGQPHGGGWGQPHGGGWGQPHGGGWSQGGGTHNQWNKPSKPKTNL
PRIO_SHEEP  ------GWGQPHGGGWGQPHGGGWGQPHGGGGWGQGGSHSQWNKPSKPKTNM
BETA_AMYLO  ----------------------DAEFRHDSGYEVHH-----------------------Q
                                   *     *

PRIO_BOVIN  KHVAGAAAAGAVVGGLGGYMLGSAMSRPLIHFGSDYEDRYYRENMHRYPNQVYYRPVDQY
PRIO_HUMAN  KHMAGAAAAGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEY
PRIO_MESAU  KHMAGAAAAGAVVGGLGGYMLGSAMSRPMMHFGNDWEDRYYRENMNRYPNQVYYRPVDQY
PRIO_MOUSE  KHVAGAAAAGAVVGGLGGYMLGSAMSRPMIHFGNDWEDRYYRENMYRYPNQVYYRPVDQY
PRIO_RAT    KHVAGAAAAGAVVGGLGGYMLGSAMSRPMLHFGNDWEDRYYRENMYRYPNQVYYRPVDQY
PRIO_SHEEP  KHVAGAAAAGAVVGGLGGYMLGSAMSRPLIHFGNDYEDRYYRENMYRYPNQVYYRPVDRY
BETA_AMYLO  KLVFFAEDVGSNKGAIIGLMVGGVV-----------------------------------
            * .  *   *.  * . * *.*   .

PRIO_BOVIN  SNQNNFVHDCVNITVKEHTVTTTTKGENFTETDIKMMERVVEQMCITQYQRESQAYYQ--
PRIO_HUMAN  SNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQ--
PRIO_MESAU  NNQNNFVHDCVNITIKQHTVTTTTKGENFTETDIKIMERVVEQMCTTQYQKESQAYYDGR
PRIO_MOUSE  SNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCVTQYQKESQAYYDGR
PRIO_RAT    SNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCVTQYQKESQAYYDGR
PRIO_SHEEP  SNQNNFVHDCVNITVKQHTVTTTTKGENFTETDIKIMERVVEQMCITQYQRESQAYYQ--
BETA_AMYLO  ------------------------------------------------------------

PRIO_BOVIN  RGASVILFSSPPVILLISFLIFLIVG
PRIO_HUMAN  RGSSMVLFSSPPVILLISFLIFLIVG
PRIO_MESAU  RSS-AVLFSSPPVILLISFLIFLMVG
PRIO_MOUSE  RSSSTVLFSSPPVILLISFLIFLIVG
PRIO_RAT    RSS-AVLFSSPPVILLISFLIFLIVG
PRIO_SHEEP  RGASVILFSSPPVILLISFLIFLIVG
```

Sequenzvergleich zwischen verschiedenen Prion-Proteinen (PRIO_) und dem menschlichen Amyloid A4 (BETA_AMYLO) der Alzheimer'schen Krankheit. Identische Aminosäuren sind durch Sterne und konservierte Aminosäuren durch Punkte gekennzeichnet. Ein grauer Balken kennzeichnet den Teil des Prion-Proteins, für den eine dreidimensionale Struktur bekannt ist.

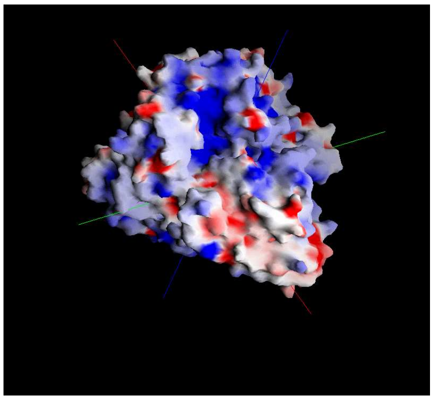
Prozessierung und Auswertung von NMR-Daten

Um Daten, die mit Hilfe der Kernresonanz (NMR)-Spektroskopie bestimmt wurden, zu verarbeiten, ist ein Einsatz von Computern unbedingt erforderlich. Insbesondere die folgenden Schritte erfordern den Einsatz von leistungsfähigen Computern:
- Reale und komplexe Fouriertransformationen
- Hilberttranformationen für den Realteil der Daten
- Einsatz der 'maximum entropy Methode (MEM) zur Verbesserung der Qualität der Spektren
- Phasenkorrekturen der prozessierten Spektren
- Deconvolution von Spektren zum Entfernen unerwünschter Lösungsmittelsignale
- Einsatz digitaler Filter zum reduzieren der Größe der Spektren
- Reale und komplexe 'linear prediction' zur Verbesserung des Signal-Rausch-Verhältnisses in den Spektren
- Einsatz von Algorithmen zur Basislinienkorrektur

Insbesondere die Fouriertranformation, die die zeitaufwendigste der oben aufgeführten Operationen darstellt, bietet die Möglichkeit zur Parallelisierung. Die dadurch zu erreichende Zeitersparnis spielt dabei eine wichtige Rolle, da die meisten Prozessierungen interaktiv durchgeführt werden.



Darstellung eines dreidimensionalen NMR-Spektrums des Birkenpollenallergens Bet v1 (links oben). Die anderen Ausschnitte zeigen die drei orthogonalen Ebenen dieses Spektrums. Die Resonanzen im Spektrum sind in orange eingezeichnet.



Strukturmodell der Adenylosuccinat-Synthetase aus *Saccharomyces cerevisiae*. Das Modell wurde auf der Basis des homologen Enzyms aus *Escherichia coli* berechnet und zeigt die elektrostatischen Eigenschaften der Proteinoberfläche. Positive und negative Ladungen sind in blau bzw. rot dargestellt.

Publikationen zum Thema

Baumann, B., Sticht, H., Sutter, M., Haehnel, W. & Rösch, P. (1996) Structure of *Synechococcus elongatus* [Fe2-S2] Ferredoxin in Solution. *Biochemistry 35*, 12831-12841.

Beißinger, M., Sticht, H., Sutter, M., Ejchart, A, Haehnel, W. & Rösch, P. (1997) The solution structure of Cytochrome C6 from the thermophilic *Synechococcus elongatus, Structure*, zur Veröffentlichung eingereicht

Seidel, G., Adermann, K., Schindler, T., Ejchart, A., Jaenicke, R., Forssmann, W.G. & Rösch, P. (1997) Solution Structure of PDIP: A homolog of the shortsighted gene product, *J. Biol. Chem..*, zur Veröffentlichung eingereicht

Sticht, H., Willbold, D. & Rösch, P. (1994) Molecular Dynamics Simulation of Equine Infectious Anemia Virus Tat Protein in Water and in 40% Trifluoroethanol, *J. Biomol. Str. Dyn. 12*, 19-36.

Sticht, H., Wildegger, G., Bentrop, D., Darimont, B., Sterner, R. & Rösch, P. (1996) An NMR-derived model for the solution structure of oxidized *Thermotoga maritima* 1[Fe4-S4] Ferredoxin, *Eur. J. Biochem. 237*, 726-735.

Sticht, H., Gallert, K.-C., Krauss, G & Rösch, P. (1997) Homology Modelling of Adenylosuccinate Synthetase from *Saccharomyces cerevisae* Reveals a Possible Binding Region for Single-Stranded ARS Sequences, *J. Biomol. Str. Dyn.,* Jun 97

Sticht, H., Pickford, A.R., Potts, J.R. & Campbell, I.D. (1997) Solution structure of the second type-two module of fibronectin, *J. Mol. Biol.*, zur Veröffentlichung eingereicht

Willbold, D., Rosin-Arbesfeld, R., Sticht, H., Frank, R. & Rösch, P. (1994) Structure of the Equine Infectious Anemia Virus Tat Protein, *Science 264*, 1584-1587.