Modelle für die Identifizierung der Konformationszustände von Proteinen und für die Berechnung ihrer Infrarotspektren

Verena Schultheis

München 2008

Modelle für die Identifizierung der Konformationszustände von Proteinen und für die Berechnung ihrer Infrarotspektren

Verena Schultheis

Dissertation
an der Fakultät für Physik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Verena Schultheis
aus München

München, im März 2008

Erstgutachter: Prof. Paul Tavan
Zweitgutachter: Prof. Erwin Frey
Tag der mündlichen Prüfung: 07.07.2008

Zusammenfassung

Proteine sind komplexe dynamische Systeme, denn sie weisen, neben schnellen thermischen Fluktuationen der Atome um ihre Ruhelage, grobskalige niedrigfrequente Übergänge zwischen einigen wenigen metastabilen Zuständen auf, den sogenannten Konformationen. Diese Konformationsdynamik und die dreidimensionale Struktur eines Proteins bestimmen seine biologische Funktion. Um die Konformationsdynamik zu untersuchen, kommen Computersimulationen zum Einsatz, welche ausgedehnte Trajektorien der Atomkoordinaten liefern. Aus diesen umfangreichen, hochdimensionalen Datensätzen Modelle der Konformationsdynamik zu gewinnen, ist eine schwierige mathematische und statistische Herausforderung.

Im ersten Teil dieser Arbeit wird ein Verfahren zur Konformationsanalyse von simulierten Trajektorien vorgestellt, das nicht nur strukturelle Eigenschaften der Simulationsdaten berücksichtigt, sondern die Zustände der zugrunde liegenden Dynamik anhand ihrer Lebensdauer hierarchisch klassifiziert. Der Datensatz wird dabei durch eine unscharfe Partition aus R univariaten Normalverteilungen zerlegt und eine R-dimensionale Übergangsmatrix für diese Zerlegung des Datensatzes bestimmt. Die Analyse dieser Markovmatrix beruht entweder auf einer nicht-linearen Dynamik oder auf einem iterativen Algorithmus, der diejenigen Markovzustände, die durch die jeweils schnellsten Übergänge verbunden sind, vereinigt. In beiden Fällen lässt sich eine Hierarchie von Markovmodellen aufstellen und anhand der Zeitskala der schnellsten Übergänge lassen sich plausible Modelle der Konformationsdynamik auswählen.

Experimentell dienen häufig Infrarotspektren zur Analyse der Proteinstruktur. Die dabei für Proteine charakteristischen Amidbanden entstehen durch gekoppelte Schwingungen der Atome des Proteinrückgrats. Daher enthält das Infrarotspektrum eines Proteins Informationen über seine Struktur, die jedoch aufgrund der Komplexität der Moleküle schwer zugänglich ist. Deswegen kommen häufig empirische Regeln zum Einsatz, die aus der Lage der Amidbanden Rückschlüsse auf den Gehalt an Sekundärstrukturmotiven ziehen. Stattdessen wäre es wünschenswert, Proteinspektren mit Hilfe fundamentaler Theorie zu verstehen. Quantenmechanische Methoden, wie die Dichtefunktionaltheorie, können zwar Infrarotspektren kleiner Moleküle berechnen, scheitern aber für Proteine in Lösung am Rechenaufwand. Für derartige Systeme kommen klassische Rechnungen in Frage, die als Näherung für die quantenmechanischen Phänomene empirische Kraftfelder verwenden. Die bisherigen Kraftfelder vernachlässigen die für Infrarotspektren entscheidenden Polarisationseffekte auf die Kraftfeldparameter und können daher Lösungsmitteleffekte nicht erfassen. Im zweiten Teil dieser Arbeit wird die Entwicklung eines aus dieser Kritik hervorgegangenen polarisierbaren Kraftfelds für das Proteinrückgrat beschrieben, bei dem die Parameter des Kraftfelds vom anliegenden elektrischen Feld abhängen.

Inhaltsverzeichnis

1 Einführung

Bei fast jedem Prozess in lebenden Organismen spielen Proteine eine wichtige Rolle. Als Antikörper bilden sie einen wichtigen Bestandteil der Immunabwehr. Collagen sorgt für die Stabilität der Zähne und Knochen. Das Bakterium *Halobacterium salinarum* pumpt mit Hilfe des Membranproteins Bacteriorhodopsin unter Verwendung von Lichtenergie Protonen zwischen Zellplasma und Umgebung [1]. Das ebenfalls bei Bakterien auftretende Porin befördert ganze Moleküle durch die Zellmembran [2]. Des Weiteren sind auch Enzyme Proteine, die als Biokatalysatoren lebenswichtige Prozesse bei Körpertemperatur erst ermöglichen und die den Ablauf dieser Reaktionen um ein Vielfaches beschleunigen. In den Biowissenschaften sind daher Proteine und ihre Funktionsweise ein zentraler Gegenstand aktueller Forschung. Welche Funktion allerdings ein bestimmtes Protein erfüllen kann, hängt von seiner dreidimensionalen Struktur und seiner Dynamik ab.

Fehler bei der Ausbildung der dreidimensionalen Struktur können verheerende Folgen haben. So spielen bei vielen neurodegenerativen Krankheiten, wie Alzheimer [3], Creutzfeldt-Jakob [4, 5], Huntington [6] oder Parkinson [7], Proteine eine entscheidende Rolle. Ein gemeinsames pathologisches Merkmal dieser Krankheiten ist die Aggregation fehlgefalteter Proteine im Gehirn [8–11].

1.1 Proteine

Proteine und Peptide sind kettenförmige, aus zwanzig natürlichen Aminosäuren aufgebaute Makromoleküle. Als Proteine bezeichnet man Aminosäurenketten mit einer Länge von einigen zehn bis einigen tausend Aminosäuren, während kürzere Ketten Peptide genannt werden. Die Abfolge der Aminosäuren, die sogenannte Primärstrukur des Proteins, ist bis auf eventuelle posttranslationale Modifikationen[1] im ge-

[1] Die Proteinbiosynthese erfolgt in zwei Schritten: Bei der Transkription im Zellkern wird die Erbinformation auf eine Boten-Ribonukleinsäure (engl. *messenger ribonucleic acid*, mRNA) übertragen, die bei der Translation an den Ribosomen in die Aminosäuresequenz übersetzt wird. Danach kann es zu sogenannten posttranslationalen Modifikationen kommen, die die Primärstruktur des Proteins verändern [2].

netischen Code festgelegt [12]. Sie bestimmt unter anderem die Größe, die Form, die Ladung, die Löslichkeit, die Stabilität, die Wasserstoffbindungsfähigkeit und die chemische Reaktivität des Proteins. Die Kenntnis der Primärstruktur allein erlaubt jedoch keine sicheren Vorhersagen über die räumliche Struktur eines Proteins.

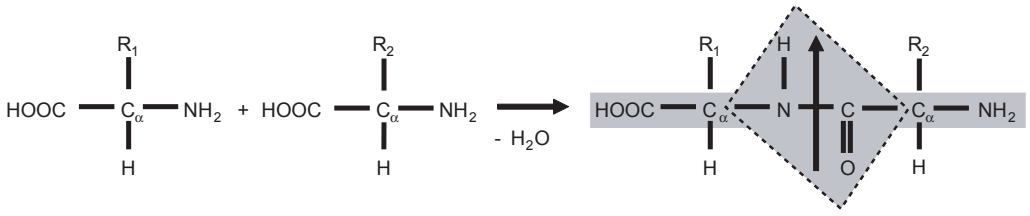

Abbildung 1.1: Kondensation zweier Aminosäuren zu einem Dipeptid. Die Atome des Peptidplättchens (gestrichelt) bilden eine verhältnismäßig starre Ebene mit einem großen Dipolmoment (Pfeil). Abbildung modifiziert nach [13].

Abbildung 1.1 zeigt links zwei Aminosäuren. Diese Moleküle bestehen jeweils aus einem zentralen Kohlenstoffatom C_α, an das eine Aminogruppe NH_2, eine Carboxylgruppe COOH, ein Wasserstoffatom H und ein Aminosäurerest R_i kovalent gebunden sind. Die verschiedenen Aminosäuren unterscheiden sich lediglich durch ihren Aminosäurerest R_i. Bei der einfachsten Aminosäure beispielsweise, dem Glycin, besteht der Rest aus einem einzigen Wasserstoffatom, bei Serin, von dem später noch die Rede sein wird, lautet er CH_2-OH.

In Abbildung 1.1 ist die Kondensationsreaktion zweier Aminosäuren zu einem Dipeptid dargestellt. Hierbei bildet das Kohlenstoffatom C der Carboxylgruppe einer Aminosäure mit dem Stickstoffatom N der benachbarten Aminogruppe unter Abspaltung eines Wassermoleküls eine Peptidbindung aus [14]. Im Dipeptid (Abbildung 1.1 rechts) ist die dadurch entstandene Peptidgruppe als Parallelogramm eingezeichnet. Sie besteht aus den beiden C_α-Atomen und den Atomen OCNH. Das π-Elektronensystem der Atome OCNH sorgt für einen partiellen Doppelbindungscharakter der CN- und CO-Bindungen. Diese Bindungen sind daher relativ torsionssteif und die Atome der Peptidgruppe bilden in diesen Kettenmolekülen eine starre Ebene, das sogenannte Peptidplättchen. Dieses weist ein großes Dipolmoment auf, das in Abbildung 1.1 durch einen Pfeil angedeutet ist.

Durch weitere Kondensationsreaktionen können sich lange Proteine bilden. Die Atome der Peptidgruppen bilden das Proteinrückgrat, das in Abbildung 1.1 grau unterlegt ist. Die starren Peptidplättchen des Proteinrückgrats sind gegeneinander um die NC_α-Bindung und um die C_αC-Bindung drehbar. Die Angabe der Torsionswin-

kel ϕ und ψ um diese Bindungen, auch Diederwinkel genannt, beschreibt daher die räumliche Struktur des gesamten Proteinrückgrats. Üblicherweise werden die Peptidgruppen entlang des Proteinrückgrats beginnend mit dem N-Terminus, das heißt mit dem Aminoende, durchnummeriert.

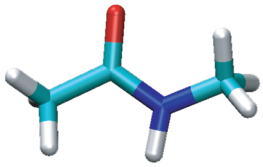

Abbildung 1.2: Das Molekül N-Methylacetamid (NMA) umfasst ein vollständiges Peptidplättchen aus Wasserstoff (weiß), Kohlenstoff (cyan), Sauerstoff (rot) und Stickstoff (blau), das an beiden Enden mit Methylgruppen abgeschlossen ist.

Abbildung 1.2 zeigt das Molekül N-Methylacetamid (NMA) in der *trans*-Konfiguration, bei der die CO- und die NH-Bindung auf entgegengesetzten Seiten der CN-Bindung liegen. Dieses Molekül kommt auch in der *cis*-Konfiguration vor, bei der die CO- und NH-Bindung in eine Richtung zeigen. Die *trans*-Konfiguration tritt bei Peptiden jedoch deutlich häufiger als die *cis*-Konfiguration auf, weil die *trans*-Konfiguration energetisch günstiger ist [2]. NMA dient häufig [15–23] als Modellmolekül für Peptide, weil es trotz seiner geringen Größe ein gesamtes Peptidplättchen umfasst (vgl. Abbildung 1.1).

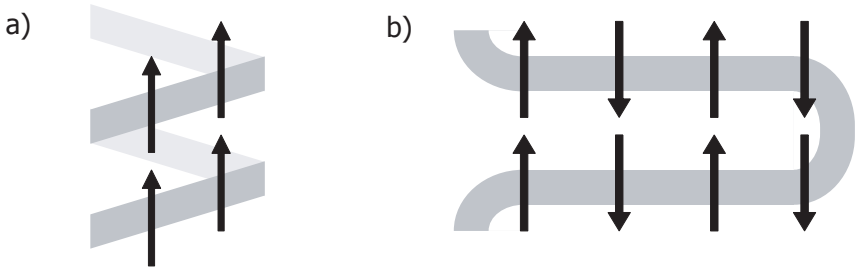

Abbildung 1.3: Zwei bei Proteinen häufige Sekundärstrukturen: Bei der α-Helix (a) addieren sich die Dipolmomente (Pfeile) der Peptidplättchen zu einem globalen Dipolmoment, während sie sich im β-Faltblatt (b) aufheben. Wie in Abbildung 1.1 ist das Proteinrückgrat grau hinterlegt. Abbildung modifiziert nach [13].

Die Anordnung nahe benachbarter Aminogruppen wird als Sekundärstruktur bezeichnet [14]. Abbildung 1.3 zeigt zwei häufige Sekundärstrukturmotive: Bei der

α-Helix (Abbildung 1.3a) ist das Proteinrückgrat spiralförmig gewunden. Die Dipolmomente der Peptidplättchen zeigen alle in eine Richtung und sorgen so für den Makrodipol der α-Helix. Die Aminosäurereste zeigen dabei nach außen. Viele, doch nicht alle Membranproteine enthalten α-Helices. Das bereits erwähnte Porin ist aus β-Faltblatt-Strukturen aufgebaut. Im β-Faltblatt (Abbildung 1.3b) windet sich das Proteinrückgrat in einer Ebene. Hier zeigen die Dipolmomente (Pfeile) benachbarter Peptidplättchen in entgegengesetzte Richtungen. Daher hat ein β-Faltblatt, anders als eine α-Helix, kein makroskopisches Dipolmoment.

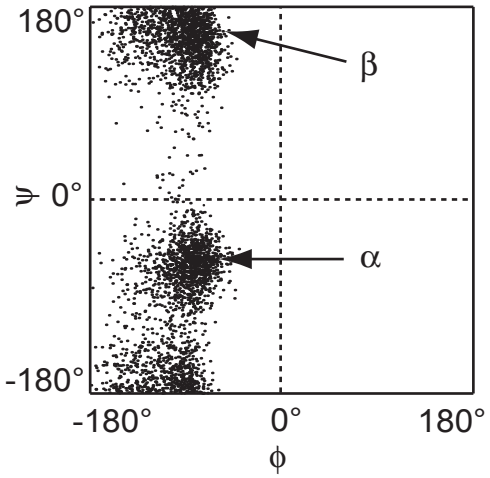

Abbildung 1.4: Die zyklischen Torsionswinkel ϕ (um die NC_α-Bindung) und ψ (um die $C_\alpha C$-Bindung) am zentralen C_α-Atom eines simulierten Serin-Tripeptids (Details zu diesem Datensatz sind in Referenz [24] enthalten). Die Punkte verteilen sich auf zwei Winkelbereiche, die α-helikalen (α) und β-Faltblatt-artigen (β) Strukturen entsprechen.

Abbildung 1.4 zeigt die bereits angesprochenen Torsionswinkel für das zentrale C_α-Atom eines Serin-Tripeptids als sogenannten Ramachandran-Plot [25] (der dieser Abbildung zugrunde liegende zirkuläre Datensatz wird in [24] ausführlich beschrieben). In dieser Darstellung kann man zwei getrennte, für die Sekundärstrukturmotive α-Helix und β-Faltblatt charakteristische Winkelbereiche unterscheiden. Andere, sogenannte *sterisch verbotene* Winkelbereiche treten nicht auf, da den entsprechenden Winkelbereichen Kollisionen zwischen mehreren Atomen entsprechen [14].

Nach dem Levinthalschen Paradoxon [26] reicht das gesamte Alter des Universums nicht aus, um alle theoretisch möglichen Proteinstrukturen statistisch auszuprobieren, selbst wenn man von einem vereinfachten Protein-Modell ausgeht, bei dem unrealistisch schnelle Übergänge zwischen nur drei möglichen Konformationen für jedes

Peptidplättchen stattfinden. In der Realität spielt sich Proteinfaltung in Bruchteilen einer Sekunde bis zu Zeiträumen von einigen Minuten ab. Es muss daher Prozesse geben, die die Proteinfaltung organisieren und beschleunigen. Der Prozess der Proteinfaltung, in dem ein Protein ausgehend von seiner Primärstruktur seine dreidimensionale Form einnimmt, die sogenannte Tertiärstruktur, ist bis heute Gegenstand der Forschung [27, 28].

1.2 Experimentelle Verfahren zur Strukturanalyse von Proteinen

Um ein detailliertes physikalisch-chemisches Verständnis der Funktionsweise von Proteinen zu erlangen, muss man die Proteinstruktur in atomarer Auflösung verstehen. Dazu gibt es zwei Methoden: die Röntgenkristallographie [29, 30] und die Kernspinresonanz (engl. *Nuclear Magnetic Resonance*, NMR) [31].

Um die biologische Funktion von Proteinen zu untersuchen, sind diese Methoden jedoch nur bedingt geeignet: Die Röntgenuntersuchungen sind in der Regel nur an Kristallen möglich, während biologische Prozesse meist in Lösung stattfinden. Selbst wenn sich Proteine kristallisieren lassen [32, 33], heißt das nicht zwangsläufig, dass sie ihre dreidimensionale Struktur und ihre biologische Funktion dabei behalten. Außerdem sind Wasserstoffatome aufgrund ihres kleinen Streuquerschnitts mit Röntgenstrahlen nicht auflösbar. Deshalb ist der Protonierungszustand von Proteinen mit dieser Methode nicht bestimmbar. Für das Verständnis vieler biologischer Prozesse, wie beispielsweise der angesprochenen Protonentransferprozesse bei Bacteriorhodopsin, ist der Protonierungszustand jedoch entscheidend.

Die NMR-Spektroskopie ist auf kleine Moleküle in Lösung beschränkt, wodurch sie für Untersuchungen an vielen Proteinen ausscheidet. NMR-Spektren nutzen Dipol-Kopplungen zwischen den Kernspins von Atomen, um daraus Informationen über die Abstände zwischen den lokalen Dipolen zu bestimmen [34]. Aus diesen Beschränkungen (engl. *restraints*) für die Abstände zwischen den Dipolen können bestimmte Algorithmen mögliche Proteinstrukturen berechnen [34]. Dabei kommen auch sogenannte Molekülmechanik-Simulationen zum Einsatz, von denen später noch die Rede sein wird. Bei großen Molekülen wird die Zuordnung durch die enorme Anzahl von möglichen Kernspinpaaren erschwert und ist ohne zusätzliche Messdaten häufig unmöglich.

Die Fourier-Transformations-IR-(FTIR-)Spektroskopie [35] verwendet ein Michelson-Interferometer, um ein Interferogramm zu erzeugen, aus dem man durch Fourier-

Transformation das gewünschte Spektrum erhält. Die zeitaufgelöste IR-Spektroskopie verwendet einen Anreg- und einen zeitverzögerten Abtastimpuls. Dadurch macht die zeitaufgelöste IR-Spektroskopie die Proteindynamik auf einer Zeitskala von Pikosekunden zugänglich.

Typische IR-Spektren von Proteinen weisen charakteristische Signaturen auf, die sogenannten Amidbanden, die durch Schwingungen der Atome in den Peptidplättchen entstehen. Die sogenannte Amid-I Mode (1600-1690 cm^{-1}) besteht im Wesentlichen in einer Streckschwingung der CO-Bindung. Die Amid-II Mode (1480-1575 cm^{-1}) und die Amid-III Mode (1229-1301 cm^{-1}) sind Kombinationen aus CN-Streckschwingung und NH-Biegeschwingung in der Ebene des Peptidplättchens (eine Beschreibung dieser und weiterer Amidmoden enthalten beispielsweise [36, 37]). Über Wechselwirkungen zwischen den Dipolen der einzelnen Peptidplättchen sind die Amidmoden verschiedener Peptidplättchen miteinander gekoppelt. Dabei spielt die Orientierung der Peptidplättchen und damit der Dipolmomente eine wichtige Rolle. Nur Molekülschwingungen, die das Dipolmoment eines Moleküls ändern, sind im IR-Spektrum sichtbar.

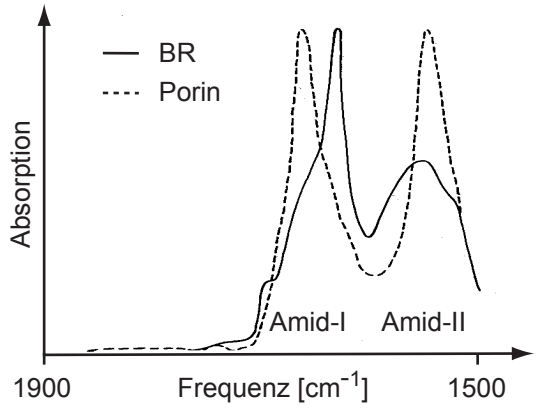

Abbildung 1.5: Experimentelle Infrarotspektren von vorwiegend α-helikalem Bacteriorhodopsin (BR) und β-Faltblatt-reichem Porin. Abbildung modifiziert nach [38].

Abbildung 1.5 zeigt die IR-Spektren der bereits erwähnten Proteine Bacteriorhodopsin und Porin. Die Amid-I Bande von Bacteriorhodopsin liegt mit 1662 cm^{-1} deutlich blauverschoben gegenüber der von Porin mit 1631 cm^{-1} [38]. Auch die Form der Amid-II Bande dieser beiden Proteine unterscheidet sich deutlich. Allgemein gilt, dass sich bestimmte Sekundärstrukturelemente, wie α-Helices oder β-Faltblätter, durch die Lage und die spektrale Form der Amidbanden auszeichnen [39]. Dieser empirische Zusammenhang findet oft bei der Interpretation von IR-Spektren Verwen-

dung, denn Form und Lage der Amidbanden kodieren die Struktur des untersuchten Proteins [40].

Auch die Lösungsmittel-Umgebung sorgt über Polarisationseffekte für eine Verschiebung der einzelnen Amidfrequenzen und damit für eine Verbreiterung der Banden. So verschiebt sich beim Molekül NMA (vgl. Abschnitt 1.1) die Amid-I Bande vom Vakuumwert 1730 cm^{-1} auf 1625 cm^{-1} in Wasser [41–43].

Bei in Wasser gelösten Proteinen erschwert das Hintergrundsignal durch das umgebende Wasser die Interpretation der IR-Spektren: Wasser hat ein Absorptionsmaximum im Bereich von 1640 bis 1650 cm^{-1} [37], das heißt im Bereich der Amid-I Bande. Daher greifen viele Experimentatoren auf andere Lösungsmittel, wie zum Beispiel D_2O zurück, deren Absorption im Bereich der Amidbanden geringer ist. Außerdem ziehen sie meist das Hintergrundsignal mittels Referenzspektren vom gemessenen Proteinspektrum ab [37, 44].

Die in diesem Abschnitt beschriebenen Methoden dienen der experimentellen Bestimmung von IR-Spektren. Die in diesen Spektren enthaltenen Strukturinformationen können bisher durch empirische Regeln nur unzureichend entschlüsselt werden. Ein detaillierteres Verständnis der Spektren erhält man durch Rechnungen, in denen der Einfluss verschiedener Strukturelemente gezielt untersucht werden kann. Der folgende Abschnitt behandelt die für die Berechnung von IR-Spektren entwickelten Methoden.

1.3 Berechnung von Infrarotspektren

Es gibt verschiedene Verfahren IR-Spektren zu berechnen [13]. Für isolierte Moleküle ist die sogenannte Normalmodenanalyse [45] üblich: Als erstes sucht man beispielsweise mit einem Gradientenabstiegsalgorithmus [46] nach dem Minimum der potentiellen Energie des Moleküls bezüglich der kartesischen Atomkoordinaten. Ausgehend vom Energieminimum bestimmt man durch finite Differenzen die Matrix der zweiten Ableitung der Grundzustandsenergie nach den Atomkoordinaten, die sogenannte Hesse-Matrix. Nach Gewichtung der einzelnen Matrixelemente mit den entsprechenden Atommassen ergeben sich durch Diagonalisierung die Absorptionsfrequenzen des Moleküls in harmonischer Näherung als Eigenwerte der massengewichteten Hesse-Matrix und die Normalmoden als ihre Eigenvektoren. Mit höheren Ableitungen lassen sich auch anharmonische Effekte berücksichtigen. Dieses Verfahren liefert für ein N-atomiges Molekül $3N$ Normalmoden, darunter jeweils drei Translations- und Rotationsbewegungen des Moleküls.

Für Moleküle in Lösung bietet sich die instantane Normalmodenanalyse (INMA [47–50]) an: Diese Methode geht von einem Schnappschuss aus einer simulierten Trajektorie des zu untersuchenden Moleküls im Lösungsmittel aus. Hier erfolgt die Geometrieoptimierung und die Bestimmung der Hesse-Matrix bei festgehaltener Lösungsmittelumgebung. Wiederholung dieser Prozedur für verschiedene Schnappschüsse aus einer Trajektorie liefert dann zeitabhängige Spektren. Außerdem lassen sich Isotopeneffekte besonders leicht durch entsprechend geänderte Atommassen bei der Berechnung der massengewichteten Hesse-Matrix berücksichtigen. Allerdings sind im Allgemeinen die berechneten Banden verbreitert, da Motional-Narrowing-Effekte vernachlässigt werden [51].

Diese Effekte treten nicht auf, wenn IR-Spektren mittels Fouriertransformation der Zeitkorrelationsfunktion (engl. *Fourier Transform of the Time Correlation Function*, FTTCF [49, 50])

$$C(t) = \sum_{i=1}^{L} \mathbf{d}(t)\mathbf{d}(t + i\Delta t) \tag{1.1}$$

der Moleküldipolmomente $\mathbf{d}$ aus einer simulierten Trajektorie berechnet werden. Hier ist L die Länge der Trajektorie und Δt ihr Zeitschritt. Nach dem Nyquist-Theorem begrenzt die Länge der Trajektorie die Auflösung $\delta\omega = 1/L$ des FTTCF-Spektrums [46]. Die FTTCF ergibt sowohl die Intensität als auch die spektrale Lage der Molekülschwingungen. Allerdings erlaubt die FTTCF nicht, Molekülmoden für die einzelnen Banden zu bestimmen. Es bietet sich an, die FTTCF-Banden durch eine INMA-Berechnung für einen Schnappschuss der Trajektorie bestimmten Moden zuzuordnen. Ein weiterer Nachteil gegenüber der INMA besteht darin, dass die gesamte Trajektorie neu berechnet werden muss, um den Effekt von Isotopenmarkierungen auf die Spektren zu bestimmen.

Beim sogenannten *Transition-Dipole-Coupling*-(TDC-)Verfahren zur Bestimmung der Amid-I Bande in IR-Spektren, geht man davon aus, dass alle CO-Dipole mit der gleichen Frequenz schwingen [52]. In späteren Veröffentlichungen [53–55] kommen auch zwei verschiedene Schwingungsfrequenzen zum Einsatz, je nachdem, ob eine bestimmte CO-Bindung dem Lösungsmittel zugänglich ist oder nicht. Über entfernungs- und orientierungsabhängige Dipol-Dipol-Kopplungen ergeben sich Normalmoden, an denen weite Bereiche des Proteinrückgrats beteiligt sind.

Die Näherung gleicher, oder maximal zweier verschiedener Schwingungsfrequenzen für alle CO-Bindungen wurde in neueren Arbeiten [18, 19, 56] aufgegeben. Diese verwenden CO-Schwingungsfrequenzen, die vom elektrostatischen *Potential* in der Peptidgruppe abhängen. Dies entspricht allerdings nicht völlig dem physikalischen Konzept der Polarisation, demzufolge das elektrische *Feld* die Elektronenwolke

eines Moleküls beeinflusst. Daher verwenden Torii [57] und Zhuang et al. [20] feldabhängige, quantenmechanisch berechnete Frequenzverschiebungen für die Amid-I Bande. Die angesprochenen Ansätze [18, 19, 53–56, 58] beschränken sich jedoch auf die CO-Streckschwingung und damit auf die Amid-I Bande. Sie ignorieren den großen Einfluss anderer Schwingungsmoden des Peptidplättchens, die an die CO-Streckschwingung koppeln.

1.4 Simulationsmethoden

Die in Abschnitt 1.3 angesprochenen Berechnungsverfahren INMA und FTTCF gehen für die Berechnung von IR-Spektren von einer simulierten Trajektorie der Atomkoordinaten aus. In den folgenden Abschnitten werden einige Simulationsverfahren vorgestellt, die zur Berechnung derartiger Trajektorien dienen.

1.4.1 Dichtefunktionaltheorie

Walter Kohn und John Pople erhielten 1998 den Chemie-Nobelpreis für die Entwicklung [59, 60] und praktische Umsetzung [61] der Dichtefunktionaltheorie (DFT, [62]). Nach dem ersten Hohenberg-Kohn-Theorem [59] lässt sich der Grundzustand eines Moleküls mit seiner Elektronendichte vollständig beschreiben, die über ein selbstkonsistentes Verfahren mit den Kohn-Sham-Gleichungen [60] berechenbar ist. Die Wechselwirkungen zwischen Elektronen gehen in das nicht analytische Austausch-Korrelations-Funktional ein. Näherungen für dieses Funktional sind die lokale-Dichte-Näherung (engl. *local density approximation*, LDA) und das gradientenkorrigierte Verfahren (engl. *generalized gradient approximation*, GGA) [63]. Der hohe Aufwand für Rechnungen mit allen Elektronen lässt sich durch die Simulation von Valenzelektronen in einem effektiven Pseudopotential der Rumpfelektronen verringern [64].

Die DFT ist auf verhältnismäßig kleine Moleküle (bis etwa 100 Atome) beschränkt [65–67]. Proteine sind jedoch mit typischerweise Tausenden von Atomen viel größer. Zudem müssen biologische Fragestellungen mit Simulationen von Proteinen in ihrer natürlichen Umgebung, das heißt in Lösung, untersucht werden. In diesem Fall kommt zu den Proteinatomen noch eine Vielzahl an Lösungsmittelatomen hinzu und der Rechenaufwand für DFT Simulationen übersteigt die Rechenleistung gegenwärtiger Computer bei Weitem. Der folgende Abschnitt erläutert daher ein alternatives Simulationsverfahren, das auch für größere Systeme geeignet ist.

1.4.2 Molekülmechanik-Simulationen

Molekülmechanik-(MM-)Simulationen beschreiben Proteine oder Peptide in Lösung als klassische Vielteilchensysteme aus Atomen [68, 69]. Die gekoppelten Newtonschen Bewegungsgleichungen der Atome werden numerisch integriert [70]. Der für den Rechenaufwand kritische Integrationszeitschritt muss dabei so gewählt sein, dass er auch die schnellsten Molekülbewegungen erfasst. Bei Biomolekülen sind die schnellsten Bewegungen die Wasserstoff-Streckschwingungen in der Größenordnung von zehn Femtosekunden. Daher sind Integrationsschritte von etwa einer halben Femtosekunde erforderlich [45]. Der SHAKE-Algorithmus [71, 72] hält die Längen der Wasserstoffbindungen innerhalb bestimmter Toleranzen konstant und verhindert so Wasserstoff-Fluktuationen. Damit sind größere Integrationsschritte von ein bis zwei Femtosekunden möglich.

Die quantenmechanischen Effekte, welche die DFT-Rechnungen explizit berücksichtigen, werden in MM-Simulationen durch eine molekülmechanische Energiefunktion parametrisiert, die auch Kraftfeld genannt wird [73–75]. Diese Energiefunktion

$$E(\mathbf{R}) = E_{\text{nonbonded}}(\mathbf{R}) + E_{\text{bonded}}(\mathbf{R}) \tag{1.2}$$

hängt von den Orten $\mathbf{R} = \{\mathbf{r_1}, \ldots, \mathbf{r_N}\}$ der N Atome ab. Der Term $E_{\text{nonbonded}}$ beschreibt Wechselwirkungen zwischen Atomen, zwischen denen drei oder mehr kovalente Bindungen liegen, oder zwischen Atomen, die zu verschiedenen Molekülen gehören. Dieser Energieanteil setzt sich aus van-der-Waals-Beiträgen und elektrostatischen Beiträgen zusammen. Während die van-der-Waals-Wechselwirkung mit dem Abstand r zwischen zwei Atomen stark (mit r^{-6}) abfällt und ab einer gewissen Entfernung vernachlässigbar ist, steigt der Aufwand für die langreichweitige (r^{-1}) Elektrostatikberechnung selbst bei besonders geeigneten Rechenverfahren [76, 77] linear mit der Anzahl der simulierten Atome. Einfache Abschneideverfahren führen bei der Elektrostatikberechnung oftmals zu nicht-tolerierbaren Simulationsartefakten [78].

Alle Wechselwirkungen zwischen Atomen, die über höchstens drei kovalente Bindungen miteinander verbunden sind, werden durch

$$E_{\text{bonded}} = \sum_i k_i (q_i - q_i^\circ)^2 + E_{\text{Dieder}} \tag{1.3}$$

beschrieben. Die Auslenkungen von internen Koordinaten q_i wie Bindungslängen und Bindungswinkeln sind bei biologisch relevanten Temperaturen klein. Sie werden in Gleichung (1.3) durch harmonische Potentiale berücksichtigt, wobei q_i° die Ruhelagen und k_i die Kraftkonstanten sind. Der Energiebeitrag E_{Dieder} in Gleichung (1.3) beschreibt die Abhängigkeit der Energiefunktion von Torsionswinkeln ξ_i wie den

Diederwinkeln an den C_α-Atomen von Proteinen (vgl. Abschnitt 1.1). Diese Winkel haben teilweise mehrere Energieminima (beispielsweise je ein Minimum für die *trans*- und die *cis*-Konfiguration eines Peptidplättchens). Die Energieabhängigkeit solcher Freiheitsgrade wird durch eine Kosinusentwicklung

$$E_{\text{Dieder}} = \sum_{\xi_j} \sum_{n_j} k_{\xi_j,n_j} \left[1 + \cos\left(n_j \xi_j + \delta_{\xi_j,n_j}\right)\right] \tag{1.4}$$

dargestellt. Der Energiebeitrag jedes Winkels ξ_j wird dabei durch eine oder mehrere Periodizitäten n_j beschrieben, zu denen jeweils eine Kraftkonstante k_{ξ_j,n_j} und eine Phasenverschiebung δ_{ξ_j,n_j} gehört. Manchmal werden zusätzlich noch Urey-Bradley-Terme [79] eingesetzt, beispielsweise zur Beschreibung von Methylgruppen. Diese Terme beinhalten die Energieabhängigkeit des Abstands zweier Atome, die beide kovalent an ein Zentralatom gebundenen sind. Im Vergleich zur Elektrostatikberechnung ist der Rechenaufwand für E_{bonded} vernachlässigbar, da hier jedes Atom nur mit einer eng begrenzten Gruppe von anderen Atomen wechselwirkt.

In konventionellen Kraftfeldern, wie beispielsweise CHARMM [80], hängen die angesprochenen Parameter lediglich von den beteiligten Atomtypen und der Art der Bindung ab. Insbesondere verwenden diese Kraftfelder dieselben Parameter und Ladungen für jedes Peptidplättchen. Sie vernachlässigen damit Kopplungen zwischen internen Koordinaten ebenso wie die erwarteten großen Polarisationseffekte durch die Umgebung. Neuere, polarisierbare Kraftfelder [73, 81] versuchen diese Polarisationseffekte teilweise zu berücksichtigen. Die drei gebräuchlichsten Ansätze dafür sind

- nach dem Elektronegativitäts-Equalisierungs-Prinzip verschiebbare Ladungen (*fluctuating charges* [82]),
- Drude-Oszillatoren (auch *shell models* genannt [83]) und
- induzierte Dipole [84, 85].

Stern et al. [81] schlagen ein MM-Kraftfeld für Proteine vor, das die polarisationsabhängige Änderung der Dipolmomente berücksichtigt. Auch der Ansatz verschiebbarer Ladungen wurde auf Proteine angewendet [86, 87]. MM-Simulationen mit derartigen polarisierbaren Kraftfeldern passen somit die elektrostatischen Eigenschaften eines betrachteten Moleküls, welche die Wechselwirkungen in $E_{\text{nonbonded}}$ beeinflussen, dynamisch an die elektrischen Felder in der Umgebung an. Sie vernachlässigen jedoch den Einfluss dieser Felder auf die Kraftfeldparameter in E_{bonded}. Eines der Ziele dieser Arbeit ist es, ein Kraftfeld zu entwickeln, das genau diesem Effekt Rechnung trägt.

Damit berücksichtigt dieses Kraftfeld den bereits angesprochenen Einfluss des Lösungsmittels und der Proteinumgebung auf das Proteinrückgrat bei der Berechnung von IR-Spektren.

1.4.3 Hybrid-Simulationen

Während sich quantenmechanische Simulationen durch eine verhältnismäßig hohe Genauigkeit bei großem Rechenaufwand auszeichnen, bieten die weniger genauen MM-Simulationen den Vorteil deutlich kürzerer Rechenzeiten. Bei Simulationen von Proteinen in einem Lösungsmittel liegt das Hauptaugenmerk oft auf der korrekten Beschreibung der Proteineigenschaften, während bei der Beschreibung des Lösungsmittels Genauigkeitsabstriche zugunsten des Rechenaufwands möglich sind. Es bietet sich daher an, sogenannte Hybrid-Rechenverfahren [88] zu verwenden.

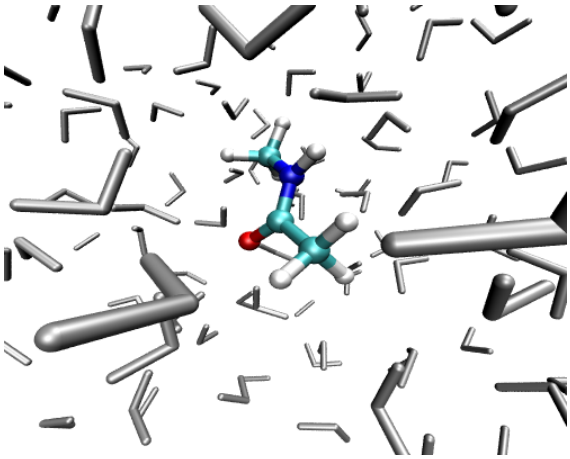

Abbildung 1.6: Bei einer QM/MM-Hybridsimulation von NMA (farbig) in Wasser (grau) wird ein Teil des Simulationssystems, das NMA-Molekül, quantenmechanisch (QM) beschrieben und der Rest klassisch mittels Molekülmechanik-(MM-)Simulationen simuliert.

Abbildung 1.6 illustriert dieses Prinzip anhand eines NMA-Moleküls in Wasser. Hier behandelt die QM/MM-Simulation das NMA-Molekül quantenmechanisch (QM), das umgebende Wasser klassisch (MM). Dabei überwiegt in der Regel der quantenmechanische Rechenaufwand bei Weitem. Verläuft die Grenze zwischen QM-Fragment und MM-Fragment durch eine kovalente Bindung, kommen Brückenatome (engl. *link atoms*) zum Einsatz [89]. Dabei wird das QM-Fragment durch ein zusätzlich eingeführtes Wasserstoffatom abgesättigt.

1.5 Konformationsanalyse

Die bisher beschriebene Technik der IR-Spektroskopie und die Methoden zur Berechnung von IR-Spektren können dazu dienen, Informationen über die dreidimensionale Struktur von Proteinen zu gewinnen. Ebenso kann man die Faltungsstruktur eines Proteins oder Peptids auch direkt anhand berechneter Trajektorien der Atomkoordinaten untersuchen.

Proteinatome unterliegen hochfrequenten, thermischen Fluktuationen um ihre Gleichgewichtslagen, wie sie auch in Festkörpern auftreten. Zusätzlich weisen Proteine grobskalige, niederfrequente Übergänge zwischen mehreren metastabilen Zuständen, den Konformationen, auf [90]. Diese langsame Konformationsdynamik ist entscheidend für die biologische Funktion. Sie ermöglicht beispielsweise einem Substrat, das aktive Zentrum eines Enzyms zu erreichen [27], oder den bereits angesprochenen Membranproteinen den Transport von Molekülen.

1.5.1 Strukturorientierte Verfahren zur Konformationsanalyse

Aus einer hochdimensionalen[1] simulierten Trajektorie geeignete vereinfachte Modelle der Konformationsdynamik zu gewinnen, ist eine komplexe Aufgabe. Diese Komplexität spiegelt sich in einer ganzen Reihe von unterschiedlichen Ansätzen wider [91–102].

Einige davon basieren darauf, dass sich Konformationen durch Minima der Energiefläche auszeichnen [27, 91, 92]. Auch die Analyse von sogenannten *recurrence plots* [103] dient zur Konformationsanalyse [93]. Die ebenfalls verwendeten Clustering-Verfahren [94, 104, 105] unterteilen einen Datensatz in – möglicherweise unscharfe (engl. *fuzzy* [106, 107]) – Teilmengen (*Cluster*) von Punkten, die einander möglichst ähnlich und von den Punkten eines anderen Clusters möglichst verschieden sind. Hierbei dienen geometrische Kriterien wie der Abstand zwischen Datenpunkten und damit strukturelle Ähnlichkeiten als Maß. Dabei ist keineswegs sicher, ob die geometrische Ähnlichkeit zur Abgrenzung verschiedener Konformationen ausreicht. Prinzipiell können verschiedene Konformationen eines Proteins im Raum der Atomkoordinaten oder Diederwinkel eng benachbart liegen, obwohl zwischen ihnen direkte Übergänge energetisch unmöglich sind. Ein rein geometrischer Clustering-Algorithmus kann diese Konformationen nicht trennen.

[1] Selbst wenn man anstelle der Atomkoordinaten eines Proteins nur die beiden Diederwinkel für jedes seiner N Peptidplättchen betrachtet, ergibt sich daraus ein Datensatz, der für jeden Zeitschritt der Trajektorie einen $2N$-dimensionalen Vektor enthält.

Ausgehend von der sogenannten invarianten Dichte eines ergodischen dynamischen Systems, nach der die Punkte der Trajektorie im Grenzwert $t \to \infty$ verteilt sind, lassen sich Konformationen wie folgt bestimmen [108]. Die Trajektorie verweilt stets eine gewisse Zeitdauer lang im Bereich einer Konformation, was zu einer hohen Dichte der Trajektorienpunkte im Bereich der Konformation führt. Daher markieren die Maxima der invarianten Dichte die Konformationen. Diese kann beispielsweise ein Gradientenaufstiegsverfahren [46] auf der invarianten Dichte auffinden. Da eine Trajektorie eine begrenzte Länge hat, steht die invariante Dichte nicht direkt zur Verfügung und muss geeignet geschätzt werden [104, 109].

Bei einer sogenannten parametrischen Dichteschätzung [109] liefert die Variation eines Parameters (wie der Standardabweichung σ der Normalverteilungen in [104]) eine einparametrische Schar von Dichteschätzungen. Schnell [110] wendet dieses Konzept auf eine Kerneldichteschätzung an [111], bei der jeder Datenpunkt das Zentrum einer univariaten Normalverteilung ist. Bei sehr kleinen Werten von σ markiert jedes Maximum der Kerneldichte einen Datenpunkt. Bei zunehmendem σ verschmelzen benachbarte Maxima nacheinander miteinander, bis schließlich ein einziges Maximum im Mittelwert der Daten übrigbleibt. Schnell [110] bestimmt für jeden Datenpunkt das zugehörige Maximum durch einen Gradientenaufstieg und teilt anhand dieser Zuordnung die Datenpunkte in disjunkte Klassen ein. Durch Variation von σ erhält er eine hierarchische Klassifikation der Daten.

Hirschberger [112] verwendet dieses Prinzip des hierarchischen Klassifikators, um mit Hilfe eines neuronalen Netzes [113, 114] Daten aus simulierten Trajektorien in Konformationen zu klassifizieren. Hillermeier et al. [115, 116] schlagen einen weiteren hierarchischen Klassifikator vor, der von einer Dichteschätzung nach dem neuronalen Kohonen-Algorithmus [117] ausgeht. Dieser Algorithmus hat jedoch den Nachteil, dass er ein lokal verzerrtes Abbild des Datensatzes liefert [118, 119]. Dies kann dazu führen, dass die Orte der Maxima der geschätzten Dichte nicht mehr mit denen der dem Datensatz zugrunde liegenden Dichte übereinstimmen. Obwohl das Verfahren von Hillermeier et al. [115, 116] völlig anders konstruiert ist als die Verfahren in [110, 112], ist es funktional diesen Gradientenaufstiegsverfahren äquivalent. Es bietet jedoch einen Ansatzpunkt, wie sich die Konformationsanalyse anhand der dynamischen Eigenschaften einer Trajektorie durchführen lässt. Dieser Ansatzpunkt und die daraus resultierenden Verfahren werden in Kapitel 2 ausführlich diskutiert. Der nächste Abschnitt gibt zunächst einen Überblick über Verfahren zur Konformationsanalyse, die sich nicht auf strukturelle Kriterien beschränken, sondern auch dynamische Informationen der Trajektorie berücksichtigen.

1.5.2 Dynamikorientierte Verfahren zur Konformationsanalyse

Grubmüller und Tavan [90] haben für ein vereinfachtes Proteinmodell gezeigt, dass diese Konformationsdynamik durch ein Markov-Modell [120, 121] beschrieben werden kann. Ein Markov-Prozess besteht aus einer endlichen Anzahl von Zuständen, zwischen denen zufällige Übergänge stattfinden. Die Wahrscheinlichkeit jedes Übergangs hängt nur vom jeweiligen Ausgangszustand ab, nicht jedoch von der weiteren Vergangenheit des Systems. Mit Hilfe einer Markov-Matrix, deren Elemente die Übergangswahrscheinlichkeiten für alle Paare von Zuständen sind, lassen sich für eine gegebene Anfangsbesetzung der Zustände eines Systems alle zukünftigen Konfigurationen dieses Systems berechnen. Viele Markov-Prozesse, insbesondere jene, die Prozesse im thermodynamischen Gleichgewicht beschreiben, genügen zusätzlich dem *Prinzip der detaillierten Bilanz* [122].

Manche Analyseverfahren [94, 95] konstruieren einen Transferoperator aus der Trajektorie, dessen Eigenwerte und Eigenvektoren Aussagen über die Stabilität metastabiler Zustände erlauben. Diesen Transferoperator bestimmen die Verfahren, indem sie den Datenraum partitionieren und die relativen Übergangshäufigkeiten zwischen jeweils zwei Partitionsvolumina des Datenraumes abzählen.

Die einfachste Partitionierung besteht darin, den Datenraum in gleich große Volumina zu unterteilen. Bei derartigen Gitterpartitionen steigt die Anzahl der Modellparameter, die statistisch aus dem Datensatz geschätzt werden müssen, exponentiell mit der Dimension des Datensatzes. Um diesen Fluch der Dimension (engl. *curse of dimensionality*) zu umgehen, beschränken Schütte et al. die Analyse auf die sogenannten essentiellen Freiheitsgrade [95, 97].

Andere Partitionierungs-Verfahren umgehen den Fluch der Dimension dadurch, dass sie die Größe der Partitionsvolumina so wählen, dass jedes Partitionsvolumen in etwa gleich viele Datenpunkte enthält. Dies ergibt sich beispielsweise bei dichteorientierten Clustering-Verfahren. Ein solches Verfahren beschreibt den Datensatz durch eine statistisch geschätzte Modelldichte [104, 109], welche ein höchstens global verzerrtes Abbild der dem Datensatz zugrunde liegenden Dichte ist. Kloppenburg und Tavan [123] gehen dazu von einer Mischung aus univariaten Normalverteilungen aus, deren Breite und Zentren sie durch einen Maximum-Likelihood-Algorithmus bestimmen [104]. Die Anzahl der Modellparameter steigt dann nicht exponentiell mit der Dimension des Datensatzes an und alle Modellparameter werden mit einer ausgewogenen Statistik geschätzt.

1.6 Ziele und Gliederung dieser Arbeit

Ein Ziel dieser Arbeit ist es, ausgehend von den in Abschnitt 1.5 beschriebenen Verfahren zur Dichteschätzung mittels einer Mischung aus univariaten Normalverteilungen Verfahren zur Extraktion von verschieden grobskaligen Markov-Modellen aus Simulationsdaten zu entwickeln. Diese Verfahren sollen in der Lage sein, nicht nur strukturelle Eigenschaften der Trajektorie zu analysieren, sondern auch dynamische Informationen der Trajektorie zu berücksichtigen. Des Weiteren sollen sie datengetrieben, also ohne Vorgabe von außen, eine Hierarchie dieser Markov-Modelle aufstellen und daraus Modelle auswählen, die den Datensatz möglichst geeignet beschreiben. Kapitel 2 beschreibt Verfahren zur statistischen Zeitreihenanalyse von Trajektorien, welche die genannten Anforderungen erfüllen. Dieses Kapitel enthält den Abdruck eines Artikels [24], den ich mit Thomas Hirschberger, Heiko Carstens und Paul Tavan verfasst habe. Die entwickelten Verfahren werden beispielhaft auf die Konformationsanalyse eines Serin-Tripeptids angewendet.

Um die Daten im Detail zu verstehen, welche die Technik der IR-Spektroskopie zur Konformationsanalyse von Proteinen oder Peptiden liefert, ist es zunächst notwendig, diese Spektren zu berechnen. Die in Abschnitt 1.3 vorgestellten Methoden INMA und FTTCF gehen dazu von simulierten Trajektorien aus. In Abschnitt 1.4 habe ich dargelegt, dass die Simulation solcher Trajektorien mit quantenmechanischen Methoden, wie der DFT, in den meisten Fällen am enormen Rechenaufwand scheitert. Die bisher in MM-Simulationen verwendeten Kraftfelder vernachlässigen wichtige Polarisationseffekte, so dass auch mit dieser Methode keine realistischen Spektren zu erwarten sind.

Kapitel 3 enthält den aus dieser Kritik hervorgegangenen Vorschlag eines polarisierbaren Kraftfelds, das zur Berechnung von Protein-IR-Spektren dient [124]. Es beschreibt die allgemeinen Prinzipien, aufgrund derer die Kraftfeldparameter bestimmt wurden und die verwendeten Rechen- und Simulationsmethoden. Weiterhin werden die Ergebnisse einer ausgedehnten DFT/MM-Simulation von NMA in Wasser und der Vergleich dieser Ergebnisse mit experimentellen Daten vorgestellt. Schließlich folgt eine Diskussion von Vakuum- und Lösungsmittel-Spektren, die mit dem polarisierbaren Kraftfeld berechnet wurden.

Das letzte Kapitel fasst die Ergebnisse dieser Arbeit zusammen und enthält einen Ausblick auf mögliche Weiterentwicklungen.

2 Extraktion von Markovmodellen der Konformationsdynamik aus Simulationsdaten

Dieses Kapitel beschreibt Verfahren zur Erstellung von Markovmodellen für die Konformationsdynamik von Peptiden anhand von simulierten Trajektorien dieser Moleküle. Es ist ein Abdruck [1] des Artikels

> Verena Schultheis, Thomas Hirschberger, Heiko Carstens, and Paul Tavan: „Extracting Markov models of peptide conformational dynamics from simulation data." *Journal of Chemical Theory and Computation* 1, 515-526 (2005),

den ich gemeinsam mit Thomas Hirschberger, Heiko Carstens und Paul Tavan verfasst habe. Zu dieser Veröffentlichung findet sich auf der Internetseite des *Journal of Chemical Theory and Computation*[2] Zusatzmaterial in Form ergänzender Grafiken, die im Anschluss an den Artikel hier ebenfalls abgedruckt sind.

[1] Reproduced with permission from The Journal of Chemical Theory and Computation, 1, 515-526, 2005. Copyright 2005 American Chemical Society.

[2] http://pubs.acs.org/journals/jctcce

JCTC Journal of Chemical Theory and Computation

Extracting Markov Models of Peptide Conformational Dynamics from Simulation Data

Verena Schultheis, Thomas Hirschberger, Heiko Carstens, and Paul Tavan*

Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität, Oettingenstrasse 67, 80538 München, Germany

Received February 3, 2005

Abstract: A high-dimensional time series obtained by simulating a complex and stochastic dynamical system (like a peptide in solution) may code an underlying multiple-state Markov process. We present a computational approach to most plausibly identify and reconstruct this process from the simulated trajectory. Using a mixture of normal distributions we first construct a maximum likelihood estimate of the point density associated with this time series and thus obtain a density-oriented partition of the data space. This discretization allows us to estimate the transfer operator as a matrix of moderate dimension at sufficient statistics. A nonlinear dynamics involving that matrix and, alternatively, a deterministic coarse-graining procedure are employed to construct respective hierarchies of Markov models, from which the model most plausibly mapping the generating stochastic process is selected by consideration of certain observables. Within both procedures the data are classified in terms of prototypical points, the conformations, marking the various Markov states. As a typical example, the approach is applied to analyze the conformational dynamics of a tripeptide in solution. The corresponding high-dimensional time series has been obtained from an extended molecular dynamics simulation.

1. Introduction

The analysis of time series[1] is important in many areas of science. Depending on the data considered, different methods are applied.[1−4] For instance, in speech recognition[5] and other fields[6] hidden Markov models[7] found important applications. They describe a dynamical system by two parametric time-discrete processes: an underlying nonobservable Markov process[8] and an observation process, defined by a sequence of conditionally independent random variables depending at each time step only on the state of the Markov chain. In many of these applications, relatively low-dimensional data are analyzed. Frequently the treatment of higher dimensional data can be simplified by first reducing the dimension, for instance using a principal component analysis.[9] Generally, the analysis of high-dimensional data mapping complex dynamical systems requires special care and the application of methods, which by construction can cope with the peculiarities of the metrics in high-dimensional data spaces.

Here, we consider a class of extremely high-dimensional and complex dynamical systems, which exhibit a largely stochastic behavior and show Markovian transitions between coarse-grained states. A typical example for such systems is the thermal motion of proteins or peptides[10] in solution. Associated time series are generated by molecular dynamics (MD) simulations[11,12] of that motion.

MD simulations treat biological macromolecules and their solvent environments as classical many-body systems composed of atoms and account for the quantum mechanical forces acting on the nuclei and caused by the electrons through a parametrized molecular mechanics force field. In MD the coupled Newtonian equations of atomic motion are integrated numerically using time steps Δt of typically 1 fs. The result of such a simulation is a trajectory $\mathbf{x}_t = \mathbf{x}(t\cdot\Delta t)$, $t = 1, 2, \ldots T$, in a high-dimensional space $\mathbf{R}^D$ (e.g. the space $\mathbf{R}^{3N}$ of the Cartesian coordinates of all $N = 100-10\,000$ atoms of a protein) describing the time sequence of configurations $\mathbf{x}_t$ sampled by the macromolecule in solution upon thermal motion. Typical simulation times are nowadays in the range of a few tens of nanoseconds ($T \approx 10^7$).

* Corresponding author phone: +49-89-2180-9220; e-mail: tavan@physik.uni-muenchen.de.

10.1021/ct050020x CCC: $30.25
Published on Web 06/08/2005

Proteins are prototypes of complex dynamical systems in soft condensed matter. In addition to high-frequency thermal fluctuations of the atoms around their equilibrium positions that are also found in solids, they show large-scale low-frequency transitions between several metastable states, the so-called conformations.[13] This slow conformational dynamics is essential for protein function in biology. Various methods[14−21] have been suggested for the extraction of protein conformations from MD trajectories. Some make use of the fact that the conformations are marked by minima of the energy landscape,[15,19,22] some apply clustering procedures based on structural similarities,[16,21] and others[14] analyze the potential energy time series by the means of recurrence plot analysis.[23]

Grubmüller and Tavan[13] have demonstrated for a simplified protein model that its conformational dynamics can be described by a simple Markov model composed of only a few conformational states. Following this principle and considering only a few so-called essential degrees of freedom Dellnitz, Schütte, and others[18,20] chose a regular lattice for discretization of the thus reduced configuration space and determined the transfer matrix of the system by counting transitions between lattice cells. They identified the conformational states defining a coarse-grained Markov model by a rather complicated analysis of the eigenvectors and -values of the transfer matrix.

Following these general concepts we here propose an alternative approach toward the analysis of high-dimensional time series, which exhibit the characteristics of a Markov chain switching among a few states. In particular, the use of a density-oriented discretization of the data space[24−26] allows us to avoid the *curse of dimensionality* inherent to grid partitions. That curse expresses the common problem, that the number of parameters, which have to be statistically estimated from the data for the construction of a simplified model, grows exponentially with the dimension of the data space.

By modifying and expanding a self-organizing and biologically plausible neural network model originally suggested for the clustering of data sets[27] but without explicitly employing the language of neural networks, we construct from the time series a transfer matrix, whose dimension is kept relatively small due to the use of the density-oriented discretization.[25] As opposed to the Kohonen algorithm[28,29] used in ref 27 for discretization, our approach does not introduce distortions into the metrics of the data space.[24,30] The analysis of the transfer matrix is either performed by a nonlinear dynamics related to the neural network used previously for clustering[27] or by a deterministic coarse-graining procedure. Both methods generate hierarchies of Markov models at varying coarseness and provide the means to identify the particular hierarchy level which most plausibly maps the generating Markov process. We start with the explanation of the methods, and, to provide a relevant example, we subsequently analyze the MD trajectory of a small peptide in water.

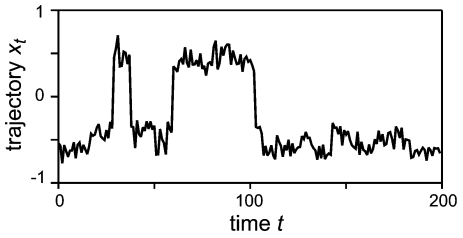

Figure 1. The first 200 steps of a time series of one-dimensional data created by a four-state Markov process. At the first glance, one can distinguish two ranges of frequent x_t values.

2. Method

For a simple graphical illustration of the employed concepts and methods, we first introduce a one-dimensional model time series, which, despite its simplicity and low-dimensionality, covers key ingredients of the problem. Figure 1 shows the first 200 steps of this time series $\mathcal{X} = \{x_t | t = 1, 2,..., T\}$, which covers $T = 10^6$ data points. The series has been generated from the Markov matrix

$$\mathbf{C}^{ex} = \begin{pmatrix} 0.8 & 0.17 & 0 & 0 \\ 0.2 & 0.8 & 0.03 & 0 \\ 0 & 0.03 & 0.8 & 0.2 \\ 0 & 0 & 0.17 & 0.8 \end{pmatrix} \tag{1}$$

by mapping the associated four-state Markov chain onto a one-dimensional dynamical system. The Markov chain generates *slow* transitions *among* the states i, $i = 1, ..., 4$. These transitions are differentiated by certain degrees of slowness: Very slow are the $2 \leftrightarrow 3$ transitions, much faster but still slow are the transitions $1 \leftrightarrow 2$ and $3 \leftrightarrow 4$. A subsequent random process completes the mapping by creating *fast* one-dimensional jumps *within* the four coarse-grained states (jumps drawn from normal distributions $g(x|x_i,\sigma)$ of standard deviation $\sigma = 0.07$ and centered at $x_i \in \{\pm 0.4, \pm 0.6\}$, see Figure 2). The resulting one-dimensional time series shares the characteristics of fast fluctuations within and differently slow transitions among coarse-grained states with peptide and protein conformational dynamics.

Note that the Markov matrix (1) generating our model time series obeys the property of detailed balance,[8,31] which requires that there are nonzero numbers f_r with

$$C^{ex}_{r'r} f_r = C^{ex}_{rr'} f_r'$$

Up to a constant factor these numbers f_r are the components $p_{r,stat}$ of the stationary distribution $\mathbf{p}_{stat} = (0.17, 0.2, 0.2, 0.17)^T/0.74$, which is the right eigenvector of $\mathbf{C}^{ex}$ to the eigenvalue $\lambda_1 = 1$. In general, R-dimensional Markov matrices $\mathbf{C}$ generating a time discrete stochastic process

$$\mathbf{p}(t + \Delta t) = \mathbf{C}\mathbf{p}(t)$$

and obeying detailed balance have a set of nice mathematical properties:[31] (i) although they are usually nonsymmetric, their eigenvalues λ_r, $r = 1,..., R$, are all real with $1 \geq \lambda_r > 0$ ($\lambda_r \geq \lambda_{r'}$ for $r < r'$), (ii) for simply connected state spaces there is exactly one largest eigenvalue $\lambda_1 = 1$ marking the

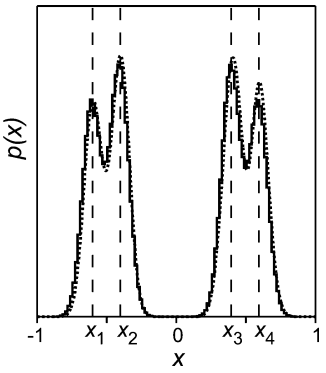

Figure 2. Histogram (solid line) and normal mixture density estimate (dotted line) of all $T = 10^6$ data points $x_t \in \mathcal{X}$ (both estimates comprise the same number $R = 100$ of local components). The four density maxima marking the Markov states are clearly distinguished. Within the two pairs (1,2) and (3,4) of the generating Gaussians considerable overlaps and between the pairs a strict separation are observed.

stationary distribution $\mathbf{p}_{stat}$, and (iii) for every initial distribution $\mathbf{p}(0)$ the iteration of the process converges to $\mathbf{p}_{stat}$.

For physical systems in thermal equilibrium the property of detailed balance frequently applies and then derives from the principle of microscopic reversibility. By applying the arguments in chapter 5.3.6b of ref 8, the following sufficient condition may be formulated for the equilibrium conformational fluctuations of a peptide sampled by an MD simulation: If the resulting trajectory $\mathcal{X}$ provides a statistically sufficient sampling of the accessible configuration space and if the observed transitions among arbitrarily defined coarse-grained states are statistically independent of the previous history of the process, i.e., are Markovian, then the associated conformational dynamics obeys detailed balance (like our simple model process does by construction).

For this simple example the particular task of time series analysis treated in this paper can now be stated as follows: Identify and reconstruct the generating Markov model (1) from the observed time series $\mathcal{X}$ as well as possible!

2.1. Partitioning the Data Space. For an ergodic system the distribution of the configurations $\mathbf{x}_t$ sampled by the trajectory $\mathcal{X}$ in the limit $t \to \infty$ defines the so-called invariant density $p_{inv}(\mathbf{x})$.[32] A parametric model for $p_{\mathrm{inv}}(\mathbf{x})$ can be estimated from a sufficiently extended sample trajectory $\mathcal{X}$ by using a mixture

$$\hat{p}(\mathbf{x}|\mathcal{W},\sigma) = \frac{1}{R}\sum_{r=1}^{R} g(\mathbf{x}|\mathbf{w}_r,\sigma) \tag{2}$$

of R univariate normal distributions $g(\mathbf{x}|\mathbf{w}_r, \sigma)$ of identical widths σ and statistical weights $1/R$ centered at points $\mathbf{w}_r \in \mathbf{R}^D$. With the exception of the number R, the model parameters, i.e., the *codebook* $\mathcal{W} \equiv \{\mathbf{w}_r | r = 1,...., R\}$ and the common width σ, are adapted to the data set $\mathcal{X}$ according to the *maximum likelihood* criterion[33] by a safely converging deterministic annealing algorithm.[24−26,34] The extraordinary robustness of this quite simple algorithm critically depends on the choice of identical widths σ for the normal distributions, although an extension toward more complicated multivariate mixture models is available.[25,26] The algorithm guarantees that the univariate normal distributions associated with the resulting optimal parameters $\mathcal{W}^{ML}$ and σ^{ML} represent roughly the same number of data points each. This property of the optimal density estimate (2) is called *load balance*[24−26] and induces a first guideline for the choice of the remaining model parameter R through the following considerations.

The components of the mixture model (2) are R class-conditional probability densities and indicate how the data belonging to class r are distributed. By Bayes' theorem every point $\mathbf{x} \in \mathbf{R}^D$ is assigned to the class r with the probability[25]

$$\hat{P}(r|\mathbf{x},\mathcal{W},\sigma) = \frac{(1/R)\, g(\mathbf{x}|\mathbf{w}_r,\sigma)}{\hat{p}(\mathbf{x}|\mathcal{W},\sigma)} \tag{3}$$

Due to the normalization

$$\sum_{r=1}^{R} \hat{P}(r|\mathbf{x},\mathcal{W},\sigma) = 1 \tag{4}$$

the probabilities (3) define a fuzzy partition of the data space when considered as functions of $\mathbf{x}$. In the limit $\sigma \to 0$ this partition becomes a crisp Voronoi tessellation[35] of the data space. Because of the load balance, each of the partition volumes covers approximately the same number T/R of data points, independently of the dimension D of the data. For a given data set of size T, the choice of the codebook size R determines T/R and thus defines the statistical quality, at which each $\mathbf{w}_r \in \mathcal{W}$ is estimated from the data $\mathbf{x}_t \in \mathcal{X}$.[26] Therefore, this type of density-oriented data space discretization can avoid the curse of dimensionality mentioned in the Introduction. For our one-dimensional example, Figure 2 compares a grid discretization (histogram) with the mixture model (2) and demonstrates the quality of the mixture estimate. Note that, because of load balance, the distribution $\tilde{p}(w)$ of codebook vectors closely models the distribution $p_{inv}(x)$ of the data ($\tilde{p} \approx p_{inv}$).[24]

2.2. Transfer Operator. The transfer operator describing the observed dynamical system is estimated using the partition described above. To simplify the notation, we extract from the trajectory $\mathcal{X}$ the set Y of all $T - 1$ pairs $y_t \equiv (\mathbf{x}_t, \mathbf{x}_{t+1})$ and define the correlation product

$$\langle f(\mathbf{x}_{t+1}) g(\mathbf{x}_t)\rangle_Y \equiv \frac{1}{T-1}\sum_{y_t \in Y} f(\mathbf{x}_{t+1}) g(\mathbf{x}_t) \tag{5}$$

where f and g are functions of $\mathbf{x}_t$. The transfer matrix $\mathbf{C}$ defined by the partition (3) then is[36]

$$C_{rr'} = \frac{\langle \hat{P}(r|\mathbf{x}_{t+1},\mathcal{W},\sigma)\hat{P}(r'|\mathbf{x}_t,\mathcal{W},\sigma)\rangle_Y}{\langle \hat{P}(r'|\mathbf{x}_t,\mathcal{W},\sigma)\rangle_Y} \tag{6}$$

Clearly, $\mathbf{C}$ depends on the parameters $\{\mathcal{W}, \sigma\}$ as well as on the choice of the codebook size R. There are R^2 matrix-elements $C_{rr'}$, which have to be statistically estimated from the $T - 1$ data points $y_t \in Y$ by evaluation of the correlation products in eq 6. To ensure sufficient statistics one should therefore demand that $R^2/T \ll 1$: This requirement thus represents a second guideline for the choice of R. Note that large values of σ, though helping to improve the statistics,

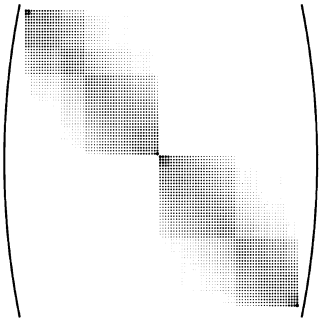

Figure 3. Transfer matrix (6) for the sample trajectory $\mathcal{X}$. The radii of the circles code the sizes of the matrix elements. On a coarse level, two diagonal blocks with non-zero elements are seen. The elements outside these blocks are substantially smaller and, therefore, are invisible here. Each of the coarse diagonal blocks decomposes into two diagonal sub-blocks, partly linked by off-diagonal blocks.

decrease the information content of **C** by smoothening. This aspect of the dependence of **C** on the fuzziness parameter σ is further discussed in section 2.7.

Because the partition functions $\hat{P}(r|\mathbf{x}, \mathscr{W}, \sigma)$ are centered around the points $\mathbf{w}_r \in \mathbf{R}^D$ and assume values close to 1 for points $\mathbf{x}$ near $\mathbf{w}_r$, the matrix (6) codes the spatial correlations between consecutive points. The elements of **C** are non-negative, and its columns are normalized to 1 ($\forall r'$: $\sum_{r=1}^{R} C_{rr'} = 1$). Therefore, **C** is an R-state Markov matrix. As it is generated from a trajectory, the associated state space is simply connected. Correspondingly, **C** has only one right eigenvector $\mathbf{p}_{stat}$ to the eigenvalue 1 marking the stationary state. As one can easily show from the definition (6) of the transfer matrix **C** and using the normalization (4) of the partition functions the stationary distribution is given by the loads of the partition volumes, i.e., $p_{r,stat} \approx \langle \hat{P}(r|\mathbf{x}_t, \mathscr{W}, \sigma)\rangle_Y$ up to corrections smaller than $1/T$. The property of load balance characteristic for our partition then implies that $\mathbf{p}_{stat}$ approximately represents a uniform distribution, that is $p_{r,stat} \approx 1/R$.

Figure 3 shows the transfer operator (6) for the time series of Figure 1. This matrix, like all other Markov matrices discussed further below, obeys detailed balance to a very good approximation: the statistical errors $|C_{rr'}p_{r',stat} - C_{r'r}p_{r,stat}|/\max\{C_{r'r}p_{r,stat}\}$ are all smaller than 1%. Because $\mathbf{p}_{stat}$ is nearly uniform, it is nearly symmetric. Apart from the eigenvalue 1, the matrix has three sizable eigenvalues (0.969, 0.442, 0.421), whereas all the remaining 96 eigenvalues are smaller than 0.002. According to refs 20 and 18 such a distribution of eigenvalues indicates the existence of two long-lived or four somewhat shorter lived metastable states. This dynamical structure of the sample trajectory is also visible in the hierarchical block structure of the depicted matrix, which clearly reveals the underlying Markov process (1). The visibility of that Markov process results from ordering the codebook elements w_r according to size ($w_r < w_{r'} \Longrightarrow r < r'$), which is only feasible in one dimension.

2.3. Analysis of the Transfer Operator. Since there is no natural ordering of the codebook vectors $\mathbf{w}_r$ in higher-dimensional cases, the analysis of transfer matrices requires other means than simple visual inspection. For this purpose

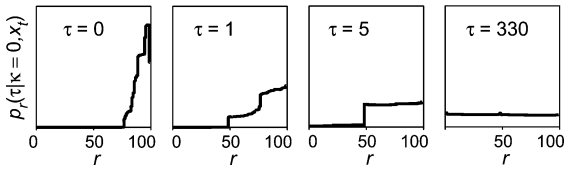

Figure 4. Linear ($\kappa = 0$) dynamics eq 8 elicited by the randomly chosen point $x_t = 0.66$ for **C** from Figure 3. Left to right: The initial distribution ($\tau = 0$) associated with x_t spreads rapidly filling predominantly the right quarter of the state space ($\tau = 1$); within the next five time steps a second metastable state appears covering predominantly the right half of the state space; at $\tau = 330$ the nearly uniform stationary distribution is reached.

we define the time-dependent probability vector[37]

$$\mathbf{p}(\tau|\mathbf{x}_t) \equiv \begin{pmatrix} \hat{P}(1,\tau|\mathbf{x}_t, \mathscr{W}, \sigma) \\ \vdots \\ \hat{P}(R,\tau|\mathbf{x}_t, \mathscr{W}, \sigma) \end{pmatrix} \tag{7}$$

whose initial components $\hat{P}(r, 0|\mathbf{x}_t, \mathscr{W}, \sigma) = \hat{P}(r|\mathbf{x}_t, \mathscr{W}, \sigma)$ are given by the posterior probabilities (3) of a given point $\mathbf{x}_t$. Furthermore we consider the evolution of the components $p_r(\tau)$ of $\mathbf{p}(\tau|\mathbf{x}_t)$ described by the following family of nonlinear differential equations

$$\frac{d}{d\tau} p_r = (\mathbf{L}\mathbf{p})_r + \kappa p_r(p_r - \mathbf{p}^2) \tag{8}$$

where the family parameter $\kappa \geq 0$ scales the nonlinear term. The matrix **L** derives from the transfer operator **C** and from the associated sampling time step $\Delta\tilde{t}$ according to

$$\mathbf{L} = \frac{1}{\Delta\tilde{t}} \ln \mathbf{C} \tag{9}$$

Note that the nonlinear dynamics (8) conserves the normalization $\sum_{r=1}^{R} p_r(\tau|\mathbf{x}_t) = 1$ of the probabilities. Since the time evolution of $\mathbf{p}(\tau|\mathbf{x}_t)$ depends on κ, we extend the notation to $\mathbf{p}(\tau|\kappa, \mathbf{x}_t)$. To calculate that evolution numerically, a discretization of (8) is used as described in Appendix A.

For an understanding of the dynamics (8), we look at the linear and the nonlinear terms of eq 8 separately. The purely linear dynamics (i.e. $\kappa = 0$) describes a Markov process of probability redistribution. Independent of the initial condition $\mathbf{x}_t$, the distribution $\mathbf{p}(\tau|\kappa = 0, \mathbf{x}_t)$ is temporarily caught in some metastable intermediate states but eventually converges toward the single stationary right eigenvector $\mathbf{p}_{stat}$ of **C**. This process is illustrated in Figure 4, which also demonstrates that $\mathbf{p}_{stat}$ is nearly uniform as claimed above.

As we explain in Appendix B, the purely nonlinear dynamics has $2^R - 1$ stationary points, each given by distributions $\mathbf{p}^{\mathscr{M}}$, which are uniform on a nonempty subset $\mathscr{M} \subset \{1,...., R\}$ and vanish elsewhere. However, only R of these distributions, the δ-distributions $p_r = \delta_{rs}$, are stable attractors of the nonlinear dynamics. The attractor δ_{rs} selected by the dynamics is defined by the largest component $p_s(0|\kappa,\mathbf{x}_t)$ of the initial distribution. Thus, the nonlinearity generates a winner-takes-all dynamics of *Darwinian selection*[27] and may be considered as the inverse of the diffusion operator.

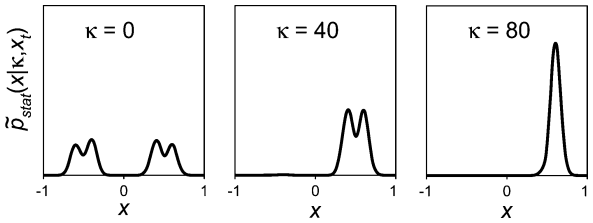

Figure 5. Stationary virtual densities associated with the starting point $x_t = 0.66$ at various strengths κ of the nonlinearity. At $\kappa = 0$ the attractor is the mixture model (2) of the invariant density depicted in Figure 2, because the generating Markov model (1) obeys detailed balance. At increasing κ the virtual density first ($\kappa = 40$) becomes confined to the two overlapping Gaussian components $i = 3$, 4 and eventually ($\kappa = 80$) to component $i = 4$ of the generating model shown in Figure 2.

By combining these mutually counteracting processes as given by eq 8 one obtains a dynamics capable of stabilizing and focusing metastable intermediates of the linear ($\kappa = 0$) relaxation process. It exhibits N_κ attractors $\mathbf{p}^n_{stat}(\kappa)$, $n = 1,...., N_\kappa$, where N_κ increases with κ ($1 \le N_\kappa \le R$). The specific attractor $\mathbf{p}^n_{stat}(\kappa)$ selected by the dynamics depends on the initial condition $\mathbf{x}_t$ and, therefore, classifies these initial conditions by $n \equiv n(\mathbf{x}_t|\kappa)$. Figure 1 in the Supporting Information illustrates how larger strengths κ of the nonlinearity stabilize increasingly short-lived metastable intermediates, prevent their diffusive spreading, and correspondingly identify metastable states at a decreasing level of coarse-graining.

2.4. Virtual Density. The distributions $\mathbf{p}(\tau|\kappa, \mathbf{x}_t) \in \mathbf{R}^R$ can be mapped onto *virtual*[38] probability densities

$$\tilde{p}(\mathbf{x}|\tau,\kappa,\mathbf{x}_t) \equiv \sum_{r=1}^{R} p_r(\tau|\kappa,\mathbf{x}_t) g(\mathbf{x}|\mathbf{w}_r,\sigma) \tag{10}$$

in the data space. For a given parameter set $\mathcal{Z} = \{\mathcal{W}, \sigma, \mathbf{C}\}$ the virtual density $\tilde{p}(\mathbf{x}|\tau,\kappa,\mathbf{x}_t)$ depends on the time τ, the nonlinearity parameter κ, and the initial condition $\mathbf{x}_t$. By eq 10, the dynamics (8) of the distributions $p_r(\tau|\kappa,\mathbf{x}_t)$ is mapped onto an equivalent temporal evolution of the virtual densities. At convergence one obtains the stationary virtual density $\tilde{p}^n_{stat}(\mathbf{x}|\kappa)$, which is associated with the initial data point $\mathbf{x}_t$ by the dynamics [$n \equiv n(\mathbf{x}_t|\kappa)$]. Particularly in the linear case ($\kappa = 0$) and for a transfer matrix (6) obeying detailed balance,[31] the virtual density $\tilde{p}(\mathbf{x}|\tau,\kappa,\mathbf{x}_t)$ converges for each $\mathbf{x}_t$ toward the mixture model (2) of the invariant density (cf. Figure 2 in the Supporting Information).

For increasing values of the nonlinearity κ, Figure 5 depicts the mapping (10) of the stationary distributions $\mathbf{p}^n_{stat}(\kappa)$ (see Figure 1 of the Supporting Information) onto the corresponding stationary virtual densities $\tilde{p}^n_{stat}(x|\kappa)$. At growing nonlinearity κ these densities $\tilde{p}^n_{stat}(x|\kappa)$ are confined to increasingly narrow and short-lived substructures of our model (2) for the invariant density. Depending on the strength κ of the nonlinearity, differently coarse-grained classes $\tilde{p}^n_{stat}(x|\kappa)$ are associated with the initial condition x_t. Thus, the stationary virtual densities turn out to represent density models for the metastable states.

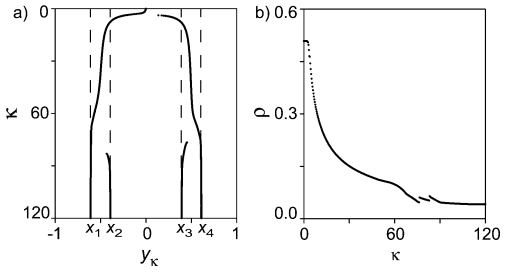

Figure 6. (a) Prototypes y_κ for $\kappa \in [0,120]$ and all data $x_t \in \mathcal{X}$. For comparison also the centers x_i for the four Gaussians associated with the states of the original Markov chain (1) are indicated by dashed lines. (b) Dependency of the average spread ρ of the stationary virtual densities on κ. The initial value $\rho \approx 0.5$ decreases almost monotonously with growing κ. At values of k near bifurcations in (a) $\rho(\kappa)$ is steeper. The small discontinuities of $\rho(\kappa)$ in the vicinity of bifurcations are due to numerics.

2.5. Moments of the Virtual Density. The first moments

$$\mathbf{y}^n_\kappa = \int \mathbf{x}\ \tilde{p}^n_{stat}(\mathbf{x}|\kappa) dx \tag{11}$$

of the stationary virtual densities (10) are obtained by integrating over the local normal distributions as[39]

$$\mathbf{y}^n_\kappa = \sum_{r=1}^{R} \mathbf{w}_r\, p^n_{r,stat}(\kappa) \tag{12}$$

Because at each κ the label n classifies the $\mathbf{x}_t$, the stationary solutions of (8) thus associate to each data point $\mathbf{x}_t \in \mathbf{R}^D$ a prototypical point $\mathbf{y}^n_\kappa \in \mathbf{R}^D$.

For our sample trajectory $\mathcal{X}$, Figure 6a depicts all prototypes y^n_κ associated with the $x_t \in \mathcal{X}$ as a function of the nonlinearity parameter $\kappa \in [0, 120]$. The figure shows that they remain invariant over wide ranges of κ while exhibiting bifurcations at certain critical values κ_i. The prototypes y^n_κ mark metastable states, characterized by fast transitions within, and slow transitions among the states of the original two-stage Markovian dynamics x_t. According to Figure 6, the boundary between *slow* and *fast* shifts toward shorter time scales with increasing κ because more and more short-lived metastable states are identified. At large κ the nonlinear dynamics eventually identifies the four prototypical points x_i characterizing the states of the original Markov model (1). Thus, the depicted bifurcation pattern reflects the hierarchical block structure of the transfer matrix (cf. Figure 3) analyzed by the nonlinear dynamics (8) at varying κ.

Higher moments of the stationary virtual densities can be calculated analogously. For a given initial condition $\mathbf{x}_t$ the variance is given by

$$\mathcal{V}[\mathbf{x}_t] = \sum_{r=1}^{R} p^{n(\mathbf{x}_t)}_{r,stat}(\kappa) [\mathbf{y}^{n(\mathbf{x}_t)}_\kappa - \mathbf{w}_r]^2 + \sigma^2 \tag{13}$$

and is—apart from the constant variance σ^2 of the Gaussians g in (10)—the sum of the squared distances between the prototypes $\mathbf{y}_\kappa$ and the codebook vectors $\mathbf{w}_r$ weighted by the probabilities (7). The value of $\rho(\mathbf{x}_t) \equiv \sqrt{\mathcal{V}[\mathbf{x}_t] - \sigma^2}$ mea-

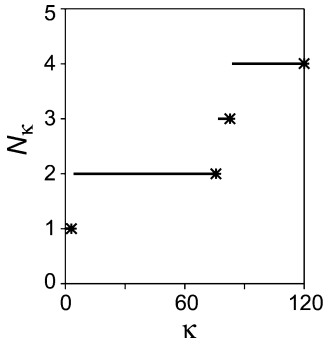

Figure 7. Number N_κ of prototypes of the trajectory from Figure 1. The asterisks mark the values of κ_ℓ, at which bifurcations occur in Figure 6a, and the value $\kappa_{max} = 120$.

sures the spread of the virtual density in data space. The dependency of the data average spread $\rho \equiv \langle \rho(\mathbf{x}_t) \rangle_Y$ on κ is plotted in Figure 6b and clearly indicates the contraction of the $\tilde{p}_{stat}$ with increasing κ.

2.6. Hierarchical Classification. For higher-dimensional dynamics one cannot visualize bifurcation patterns. Therefore other means are required to obtain insight into the hierarchy of classes identified by the nonlinear dynamics at increasing κ. A generally applicable procedure is to determine all prototypes (12) for all vectors $\mathbf{x}_t \in \mathcal{X}$, which results for each value of κ in a prototype set $\mathcal{P}_\kappa = \{\mathbf{y}_\kappa^n | n = 1,...,N_\kappa\}$ (cf. Figure 6a). The number N_κ of different stationary solutions can then be plotted as a function of κ and gives a first insight into the coarse-grained structure of the dynamics.

Figure 7 shows such a plot for our sample dynamics. The number N_κ grows monotonically with κ and remains constant within certain intervals $[\kappa_\ell, \kappa_{\ell+1}]$ ($\ell = 1, 2,...$ and $\kappa_1 < \kappa_2 < ...$). These intervals differ strongly in widths. Two large intervals belong to the values $N_\kappa = 2$ and $N_\kappa = 4$. They indicate that the system has two and four differently coarse-grained states with strongly different lifetimes. Thus, the corresponding two or four classes may be good choices for the intended construction of coarse-grained models, and we know in this case, of course, that they are. There is also a very small interval marking a three-state model, which finds no correspondence in the generating process given by eq 1. This three-state model is due to statistical fluctuations affecting the elements of the 100-dimensional transfer matrix and, therefore, the classification of the data points by the nonlinear dynamics. However, the small range of κ-values, within which the three-state model is predicted, indicates that it is not an intrinsic feature of the monitored time series. Similar structures are expected to be found in such plots whenever a reasonably clear-cut separation of time scales happens to exist in the dynamics represented by the transfer matrix (6). Also here large intervals with constant N_κ will point to plausible models.

It now remains to be seen at which values of κ these models should be determined. For this purpose we use the observation (cf. Figure 6a) that the prototypes $\mathbf{y}_\kappa^n$ do not vary much as κ approaches a critical bifurcation value κ_ℓ from below. Therefore, we reduce the continuous family $\{\mathcal{P}_\kappa | \kappa \in \mathbf{R}_0^+\}$ of prototype sets $\mathcal{P}_\kappa$ to a minimal discrete family

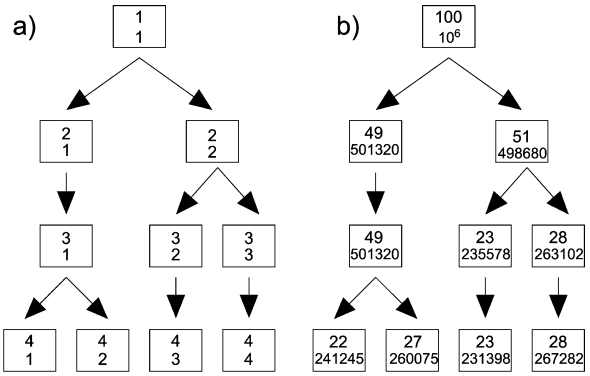

Figure 8. (a) The hierarchy of prototypes as a tree. The nodes (boxes) characterize the prototypes $\mathbf{y}_\ell^n$ by the upper index ℓ and lower index n. The edges (opposite to the arrow) denote the association of a prototype $\mathbf{y}_{\ell+1}^n$ to a prototype $\mathbf{y}_\ell^{n'}$. (b) The classification of the data set $\mathcal{X}$. Here, the upper index denotes the number of codebook vectors, $\mathbf{w}_r$ associated with the respective prototype, and the lower index the corresponding number of data $\mathbf{x}_t$.

$\{\mathcal{P}_\ell | \ell = 1,..., \ell_{max}\}$ of ℓ_{max} prototype sets $\mathcal{P}_\ell$ by selecting the prototype sets $\mathcal{P}_\ell \equiv \mathcal{P}_{\kappa_\ell}$ which are located just before the jumps of N_κ to higher values. In our simple example these values κ_ℓ are marked by asterisks in Figure 7.

In the next step we arrange the thus determined discrete family of prototype sets into a hierarchy by associating with a higher-level prototype $\mathbf{y}_\ell^{n'}$ each lower-level prototype $\mathbf{y}_{\ell+1}^n$, whose probability vector $\mathbf{p}(\tau|\kappa, \mathbf{y}_{\ell+1}^n)$ converges at $\kappa = \kappa_\ell$ under the dynamics (8) to $\mathbf{p}^{stat}(\kappa_\ell, \mathbf{y}_\ell^{n'})$. For our standard example the resulting hierarchy is drawn as a directed tree in Figure 8a.

Analogously we can classify the codebook $\mathcal{W}$ and the data set $\mathcal{X}$ on the hierarchy level ℓ by calculating for each of their elements the first moment (12) of the stationary virtual density at $\kappa = \kappa_\ell$. The result of this classification for the sample data set is shown in Figure 8b. As a result of the load balance of the mixture model (2) mentioned in 2.1, the percentage of codebook vectors $\mathbf{w}_r$ associated with a prototype $\mathbf{y}_\ell^n$ reproduces approximately the respective percentage of data points $\mathbf{x}_t$.

2.7. Extracting a Markov Model at a Hierarchy Level. Having set up a hierarchy of classifiers, which, at each hierarchy level ℓ, associate the codebook vectors $\mathbf{w}_r$ and data points $\mathbf{x}_t$ to one of N_ℓ prototypes $\mathbf{y}_\ell^n$, $n = 1,..., N_\ell$, it remains to be clarified as to how one should calculate correspondingly coarse-grained N_ℓ-state Markov matrices $\mathbf{C}^\ell = \{C_{nn'}^\ell\}$. As discussed above, such matrices can represent plausible coarse-grained descriptions of the observed dynamics, if the associated number N_ℓ of prototypes has been found to be stable over a wide range of the nonlinearity parameter κ.

There are two different choices for the computation of the $\mathbf{C}^\ell$. One can (i) reduce the original R-state Markov matrix (6) by using the classification of the codebook vectors $\mathbf{w}_r$ or (ii) directly set up the coarse-grained matrices by employing the classification of all data $\mathbf{x}_t$. To make notation simpler

we consider a single selected hierarchy level ℓ and discuss choice (i) first.

Let $I_n = \{r|\mathbf{y}_\ell(\mathbf{w}_r) = \mathbf{y}_\ell^n\}$ be the set of all indices r of codebook vectors $\mathbf{w}_r$, which are classified to a given prototype $\mathbf{y}_\ell^n$. Summing the associated partition volumes $\hat{P}(r|\mathbf{x}_t, \mathscr{W}, \sigma)$ according to

$$\tilde{P}_n(\mathbf{x}_t) \equiv \sum_{r\in I_n} \hat{P}(r|\mathbf{x}_t, \mathscr{W}, \sigma) \tag{14}$$

we obtain the partition function of the prototype $\mathbf{y}_\ell^n$, which measures the *posterior* probability that the point $\mathbf{x}_t$ belongs to $\mathbf{y}_\ell^n$. Like the original R-state Markov model defined by eq 6 also the correspondingly reduced Markov matrix $C_{nn'}^\ell$ should fulfill the analogous relation

$$C_{nn'}^\ell = \frac{\langle \tilde{P}_n(\mathbf{x}_{t+1})\, \tilde{P}_{n'}(\mathbf{x}_t)\rangle_Y}{\langle \tilde{P}_{n'}(\mathbf{x}_t)\rangle_Y} \tag{15}$$

with $n, n' \in \{1,..., N_\ell\}$. Inserting the *posterior* probabilities (14) into (15) and taking into account the definition (6) as well as the fact that the index sets I_n and $I_{n'}$ are disjoint for all classes $n \neq n'$, we obtain a reduced Markov matrix $C_{nn'}^\ell$ from the original matrix $C_{rr'}$ by

$$C_{nn'}^\ell = \frac{\sum_{r'\in I_{n'}} [\langle \hat{P}(r'|\mathbf{x}_t, \mathrm{W}, \sigma)\rangle_Y \sum_{r\in I_n} C_{rr'}]}{\sum_{r'\in I_{n'}} \langle \hat{P}(r'|\mathbf{x}_t, \mathscr{W}, \sigma)\rangle_Y} \tag{16}$$

Note that this coarse graining procedure of Markov matrices preserves detailed balance, i.e., if detailed balance holds for $C_{rr'}$, it also holds for $C_{nn'}^\ell$ as can be seen by a few lines of algebra. The stationary distribution at level ℓ follows by $p_{n,stat}^\ell = \sum_{r\in I_n} p_{r,stat}$ from $\mathbf{p}_{stat}$ associated with $\mathbf{C}$. For our synthetic sample time series, in particular, one can additionally show that in the limit of infinite sampling the detailed balance of the generating Markov matrix (1) induces detailed balance also into the discretized transfer operator $\mathbf{C}$ given by eq 6. In this case one therefore expects that detailed balance holds at all levels of coarse graining up to statistical errors.

Following choice (ii) we can alternatively count all initial data pairs $(\mathbf{x}_{t+1}, \mathbf{x}_t)$ which the nonlinear dynamics maps to the prototype pairs $(\mathbf{y}_\ell^n, \mathbf{y}_\ell^{n'})$ and all initial data points $\mathbf{x}_t$ mapped to $\mathbf{y}_\ell^{n'}$. Calling the respective numbers $T_{nn'}$ and $T_{n'}$, with $\sum_n T_{nn'} = T_{n'}$, the reduced transfer matrix is

$$\tilde{C}_{nn'}^\ell = \frac{T_{nn'}}{T_{n'}} \tag{17}$$

For overlapping coarse-grained classes both choices will overestimate the transition probabilities due to unavoidable Bayesian decision errors. For explanation consider our standard example, in which the classes associated with the Gaussians $i = 1, 2$ and $i = 3, 4$ of the generating dynamics exhibit considerable overlaps (cf. Figure 2). Even an optimal Bayesian classifier[26] will, e.g., erroneously associate data $x_t > -0.5$ that have been drawn from the normal distribution 1 to class 2. As a result, fast transitions within class 1 are erroneously counted as $1 \rightarrow 2$ transitions, and the corresponding off-diagonal element $\tilde{C}_{21}^4$ of a four-state Markov model is overestimated at the expense of the diagonal element $\tilde{C}_{11}^4$. The size of this Bayesian decision error can be estimated by comparing the four-state Markov matrix

$$\tilde{\mathbf{C}}^4 = \begin{pmatrix} 0.72 & 0.24 & 0.00 & 0.00 \\ 0.28 & 0.73 & 0.03 & 0.00 \\ 0.00 & 0.03 & 0.72 & 0.28 \\ 0.00 & 0.00 & 0.25 & 0.72 \end{pmatrix} \tag{18}$$

which has been calculated by eq 17 at the highly plausible level $\ell = 4$ of the hierarchy in Figure 8, with the generating Markov model (1). In fact, a Bayesian classification of the data $x_t \in \mathscr{X}$ (using the knowledge on the four class-conditional distributions from which the data have been drawn) numerically reproduces the four-state Markov matrix (18). As a result of the Bayesian decision error the estimated lifetimes

$$\tau_n = \frac{\Delta\tilde{t}}{1 - \tilde{C}_{nn}^\ell} \tag{19}$$

of the various coarse-grained states n are lower bounds to the true lifetimes of the generating dynamics.

For a related reason, small additional errors of this type will be introduced, if a Markov model on a given hierarchy level is estimated by the efficient reduction algorithm (16) instead by the computationally more demanding counting algorithm (17). The additional errors arising in the description of transitions among overlapping states are now due to the fuzziness σ of the partition (3) used both for the original discretization (6) of the transfer operator and for the coarse-grained partition functions (14). Correspondingly, they can be reduced by decreasing σ beyond the value σ^{ML} determined by the maximum likelihood estimate (cf. section 2.1). For our standard example and $\sigma = \sigma^{ML}$, they can be estimated by comparing the four-state matrix

$$\mathbf{C}^4 = \begin{pmatrix} 0.69 & 0.26 & 0.00 & 0.00 \\ 0.31 & 0.71 & 0.03 & 0.00 \\ 0.00 & 0.03 & 0.71 & 0.32 \\ 0.00 & 0.00 & 0.26 & 0.68 \end{pmatrix} \tag{20}$$

extracted by eq 16 with the optimal estimate (18) and the underlying Markov model (1). For instance, due to the Bayesian decision error the lifetime of state 1 is underestimated in (18) by about 30%, to which the fuzziness affecting (20) adds another 5%.

Fortunately, overlapping coarse-grained states are unlikely in high-dimensional data spaces, particularly in the ones one may use for the characterization of peptide conformational dynamics. Therefore, the unavoidable Bayesian decision errors are expected to be small. For the same reason the use of a fuzzy partitioning should not introduce large errors here, because overlapping partition volumes will mainly occur in the mapping of statistically predominant states and will then be combined by eq 14 into the associated coarse partition volumes. Because they then belong to the same state, they cannot affect the critical statistics of interstate transitions.

As a result, both algorithms should be equally applicable here and Bayesian decision errors will hardly deteriorate the results.

2.8. Alternative Construction of a Hierarchy. To check our results, we now will explain an alternative, deterministic, and very fast algorithm for constructing a hierarchy of coarse-grained Markov models from the original R-dimensional transfer matrix (6) and for identifying most plausible levels within that hierarchy. Here, the basic idea is to consecutively unite those Markov states that are mutually connected by the fastest transitions.

The alternative procedure is solely applicable to dynamical processes obeying *detailed balance* (see the introductory remarks to section 2), because this principle allows us to uniquely assign a *time scale* to the mutual transitions at the various levels of the hierarchy. The R-state Markov model (6) obeys detailed balance, if

$$C_{rr'}\langle \hat{P}(r'|\mathbf{x}_t, \mathscr{W}, \sigma)\rangle_Y = C_{r'r}\langle \hat{P}(r|\mathbf{x}_t, \mathscr{W}, \sigma)\rangle_Y \qquad (21)$$

meaning that the probability flow between any two states r and r' is equal in the stationary distribution. Dividing eq 21 by the components of the stationary distribution we immediately see that the matrix

$$D_{rr'} \equiv \frac{C_{rr'}}{\langle \hat{P}(r|\mathbf{x}_t, \mathscr{W}, \sigma)\rangle_Y} \qquad (22)$$

is symmetric. Therefore, its off-diagonal elements measure flow rates of the mutual transitions $r \leftrightarrow r'$, and we denote the maximal rate by D_{max}.

If we collect the index pair $\{r, r'\}$ belonging to D_{max} into an index set I_{R-1} and define one-member index sets I_n, $n = 1,..., R-2$, to contain the indices r'' of the remaining states, we obtain the $R-1$ index sets I_n, required for a first coarse-graining of partition volumes (14) and Markov matrices (16). The resulting $(R-1)$-state Markov matrix $\hat{\mathbf{C}}^{R-1}$ can be considered as the level $\ell = R-1$ of a model hierarchy, whose basis is formed by $\mathbf{C} \equiv \hat{\mathbf{C}}^R$. At this level, the coarse-grained partition volumes (14) provide a Bayesian classifier for the data $\mathscr{X}$ in terms of $R-1$ Markov states.

The above process of combining the fastest mixing states into new and coarser states can be iterated until the level $\ell = 2$ just below the top of the hierarchy is reached. In this recursive coarse-graining scheme, the set $\hat{\mathscr{P}}^{\ell-1}$ of prototypes $\mathbf{y}_n^{\ell-1}$ is obtained for $n = 1,..., \ell-2$ by

$$\hat{\mathbf{y}}_n^{\ell-1} = \mathbf{y}_{r''}^{\ell}, \quad \text{if } I_n = \{r''\} \qquad (23)$$

and for $n = \ell-1$ by

$$\hat{\mathbf{y}}_n^{\ell-1} = \frac{A_r^{\ell}\hat{\mathbf{y}}_r^{\ell} + A_{r'}^{\ell}\hat{\mathbf{y}}_{r'}^{\ell}}{A_r^{\ell} + A_{r'}^{\ell}}, \quad \text{if } I_n = \{r, r'\} \qquad (24)$$

from $\hat{\mathscr{P}}^{\ell}$, where the coefficient A_r^{ℓ} denotes the number of codebook vectors $\mathbf{w}_r$ previously united into the prototype $\mathbf{y}_r^{\ell}$. Note here the initial conditions $\hat{\mathscr{P}}^R = \mathscr{W}$ and $\hat{\mathbf{y}}_r^R = \mathbf{w}_r$.

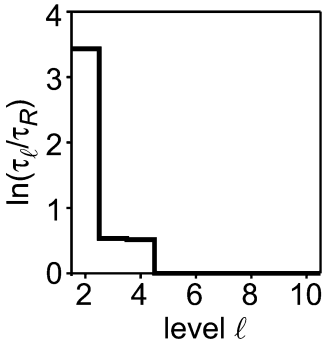

Figure 9. Time scales τ_ℓ for the last nine steps of the recursive coarse-graining procedure applied to our standard example. As time unit we have chosen the fastest time scale τ_R, which is given by the smallest eigenvalue λ_{min}^R of **C**. For explanation see the text.

At each level ℓ the fastest relaxation time scale can be characterized by considering the quantity

$$\tau_\ell \equiv \frac{1}{1-\lambda_{min}^{\ell}} \qquad (25)$$

where λ_{min}^{ℓ} is the smallest eigenvalue of $\hat{\mathbf{C}}^{\ell}$. Due to the consecutive removal of the most rapidly mixing states during our recursive coarse-graining, τ_ℓ is expected to increase in the sequence $\ell = R,..., 2$. Therefore the question, whether a given level of the resulting hierarchy furnishes a plausible coarse-grained model for the observed dynamics, can be decided by considering the ℓ-dependence of the fastest relaxation time scale τ_ℓ remaining at level ℓ. If $\tau_\ell \gg \tau_{\ell+1}$, then the model at level ℓ is considered to be plausible, because a large jump toward slower time scales indicates the presence of slowly mixing, i.e., metastable states at ℓ and of rapidly mixing states at the preceding level $\ell+1$.

Figure 9 shows such a plot for our standard example using a logarithmic time scale. Large jumps of $\ln(\tau_\ell/\tau_R)$ occur when ℓ approaches the levels four and two from above. Thus the plot clearly reveals the hierarchical four- and two-scale structure of our example. Although the model hierarchy obtained by recursive coarse-graining generally differs from that generated by the nonlinear dynamics, the two procedures predict essentially identical models at the relevant levels $\ell = 2, 4$. Here, particularly the matrices $\hat{\mathbf{C}}^{\ell}$ calculated by the recursion are identical to the $\mathbf{C}^{\ell}$ obtained by version (i) of the dynamics-based procedure. Thus, the aim of validating the latter procedure has been reached.

2.9. Merits and Deficiencies of the Various Schemes. Up to this point we have introduced three algorithmic schemes by which one can derive a hierarchy of coarse-grained Markov models $\mathbf{C}^{\ell}$ from the transfer operator **C** defined by eq 6.

Scheme 1, which represents version (ii) of the dynamics-based procedure, relies at each hierarchy level ℓ on a crisp partitioning of the data $\mathbf{x}_t \in \mathscr{X}$ into N_ℓ coarse-grained classes n by the nonlinear dynamics (8) and a subsequent counting (17) of transitions among classes. Remarkably, in this scheme the fuzziness of the partition employed for the evaluation of the transfer operator **C** does not introduce errors into the

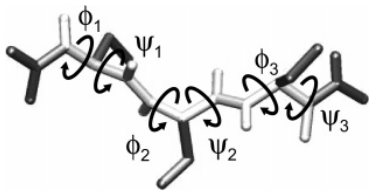

Figure 10. Tripeptide consisting of the backbone (light gray) and the serine side chains (dark gray) and definition of the dihedral angles (ϕ_i,ψ_i), $i = 1, 2, 3$. Most of the hydrogen atoms and all surrounding water molecules are omitted for clarity of representation.

computation of the coarse grained models $\mathbf{C}^{\prime}$. Experience has even shown that the nonlinear dynamics for classification of the data becomes more stable, if one increases the fuzziness of the partition in the computation of $\mathbf{C}$. In scheme 1, solely the limited statistics and Bayesian decision errors, which are unavoidable in the case of overlapping coarse-grained states, are sources of errors. The scheme is computationally expensive, because all T data points have to be classified by iteration of eq 8 at the $\prime = 1,\ldots, \prime_{max}$ stages of the hierarchy.

Scheme 2, which represents version (i) of the dynamics-based procedure, classifies solely the R codebook vectors $\mathbf{w}_r$ by the nonlinear dynamics (8) and builds the hierarchy of $\prime_{max}$ Markov models $\mathbf{C}^{\prime}$ by a corresponding coarse-graining (16) of $\mathbf{C}$. Because the original partition used for the computation of $\mathbf{C}$ is preserved, small errors may be induced by its fuzziness. Therefore, one should reduce the fuzziness of the partition for the computation of $\mathbf{C}$ in this case below the value σ^{ML} obtained from codebook optimization. The computational effort is smaller by a factor R/T than in scheme 1.

Scheme 3 directly constructs the $\mathbf{C}^{\prime}$ from $\mathbf{C}$ by a deterministic and iterative unification of partition volumes and, thus, avoids costly iterations of the nonlinear dynamics (8). It is the computationally fastest procedure, shares the fuzziness errors with scheme 2, but is applicable solely to transfer operators exhibiting detailed balance to a good approximation. Because detailed balance requires extended trajectories this requirement limits the applicability of scheme 3.

In contrast, the other two schemes can also cope with a less extensive sampling and will render reasonable Markov models for a trajectory simulating equilibrium fluctuations (or for a set of trajectories starting from a given nonequilibrium state) as long as the data exhibit Markovian transitions among the various coarse-grained states. Therefore, they are also capable of modeling nonequilibrium relaxation processes. All three schemes provide the means to distinguish relevant levels of the hierarchy from spurious ones, and in all cases the computational effort is very small as compared to the cost of generating an extended MD-trajectory for a peptide in solution.

3. Sample Application

In the above explanation of algorithms we have considered a simple one-dimensional time series for purposes of illustration. To give an impression of more realistic applications, we will now consider a six-dimensional time series obtained from a 50 ns MD simulation of a serine tripeptide in water at room temperature and ambient pressure. Details of the simulation are given in Appendix C.

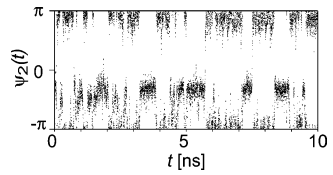

Figure 11. Time evolution of the angle ψ_2 (cf. Figure 10) during the first 10 ns of the MD simulation. Note that ψ_2 is a circular variable.

Figure 10 shows a configuration of the tripeptide molecule. At physiological temperatures its backbone (light gray) exhibits only six large-scale torsional degrees of freedom around chemical bonds, which are described by the dihedral angles ϕ_i and ψ_i. Thus, the temporal fluctuations of these six angles can be employed to determine the conformational dynamics of the backbone sampled by the MD trajectory.

Correspondingly we have generated a time series $\mathscr{X}$ from the MD trajectory, which consists of $T = 50\,001$ six-dimensional vectors $\mathbf{x}_t = [\phi_1(t),\psi_1(t),\ldots,\psi_3(t),\psi_3(t)] \in (-\pi,\pi]^6$ and represents the backbone configurations at sampling intervals $\Delta\tilde{t} = 1$ps. Note that the torsion angles are circular variables and have to be treated accordingly.[21,40]

While the ϕ-angles fluctuate around $\approx -\pi/2$ (data not shown), the ψ-angles show a more interesting behavior. As an example Figure 11 shows the angle $\psi_2(t)$ during the first 10 ns of the simulation. Two ranges of values for $\psi_2(t)$ can be distinguished. One is given by the interval $I_\alpha = [-5\pi/6, \pi/6]$ and the other by its complement I_β. The angles ψ_1 and ψ_3 exhibit a similar bimodal behavior (data not shown) as is typical for polypeptides or proteins. Following the usual nomenclature[41] we classify local backbone configurations as α-helical, if $\psi_i \in I_\alpha$, and otherwise as extended or β-strandlike. Because each ψ-angle is either in the α- or in the β-range, we a priori expect the tripeptide to populate $2^3 = 8$ different conformations.

For time series analysis we first modeled the data distribution by a 25-component mixture density $\hat{p}(\mathbf{x}_t|\mathscr{W}^{ML},\sigma^{ML})$ as described in section 2.1. Here, the value $R = 25$ was chosen, because the quotient $R^2/T \approx 1.25\%$ appeared to be small enough as to enable a reasonably accurate statistics in the estimation of the transfer operator $\mathbf{C}$ by (6). $\mathbf{C}$ turned out to have eight large eigenvalues in the range [1.0,0.78]. The remaining eigenvalues were all smaller than 0.48. As discussed in section 2.2 such a structure of the eigenvalue spectrum points toward an eight-state model in agreement with our above expectation.

The plausibility of an eight-state Markov model was subsequently confirmed by classifying the data $\mathbf{x}_t$ through the nonlinear dynamics (8) at varying κ, because $N_\kappa = 8$ prototypes $\mathbf{y}_\kappa^n$ were found to be stable attractors of that dynamics over a wide range of κ-values. A classification of the three ψ_i values of these prototypes in terms of the α- and β-ranges introduced above then revealed that the $\mathbf{y}_\kappa^n$ are

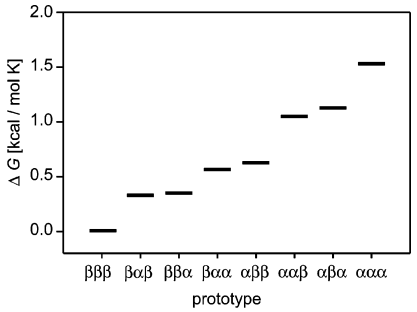

Figure 12. Free energy differences $\Delta G_n = G_n - G_1$ for an eight-state model with states n ordered according to increasing free energy $G_n = -k_B T \ln(P_n)$ and labeled by the simple (α/β)-classification of the three ψ-angles occurring in the prototypes $\mathbf{y}_\kappa^n$.

characterized by the eight possible triples $\alpha\alpha\alpha$, $\beta\alpha\alpha$, ..., which can be formed from the symbols α and β. Also the recursive coarse-graining of $\mathbf{C}$ explained in section 2.8 and a time scale analysis analogous to that in Figure 9 (see Figure 3 of the Supporting Information) indicated an eight-state model. Furthermore, the two thus determined eight-state Markov matrices turned out to be identical, i.e., $\mathbf{C}^8 = \hat{\mathbf{C}}^8$ (data not shown).

The nonlinear dynamics classification of the data $\mathbf{x}_t \in \mathcal{X}$ at the eight-state level of the model hierarchy yielded the statistical weights P_n of the states $n = 1, ..., 8$. According to the arguments in ref 13 they determine the free energies $G_n = -k_B T \ln(P_n)$ of these states, where k_B is the Boltzmann constant and T is the temperature. The resulting relative energies of the eight conformational states are depicted in Figure 12. Interestingly the fully extended conformation $\beta\beta\beta$ is seen to be energetically most favorable and, therefore, is most frequently encountered in the trajectory. Furthermore, a $\beta \rightarrow \alpha$ transition is seen to be energetically most favorable at ψ_2 and least favorable at ψ_1.

However, the dynamical connectivity of the eight states, which is visualized in Figure 13 by a plot of the Markov matrix $\mathbf{C}^8$, does not simply reflect the energetic state ordering. For instance, transitions $\beta\beta\beta \rightarrow \beta\beta\alpha$ are seen to be more likely than $\beta\beta\beta \rightarrow \beta\alpha\beta$ although the latter target state has a slightly lower free energy than the former. Furthermore, the various conformations are mainly connected by single $\beta \rightarrow \alpha$ transitions at individual angles ψ_i, whereas correlated transitions at pairs of these angles are quite rare.

By looking at further details of the connectivity displayed in Figure 13 and of the energetics shown in Figure 12, by analyzing the structures of the prototypes $\mathbf{y}_\kappa^n$ through molecular graphics, etc. one could now derive a lively picture and physical understanding concerning the conformational dynamics of serine tripeptide in water. However, these issues are beyond the scope of this paper. In the present context the given example solely serves to illustrate the kind of insights into complex dynamical systems, which now can be gained by applying the methods of time series analysis outlined above.

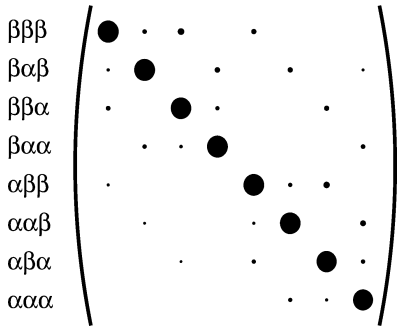

Figure 13. Graphical representation of the Markov matrix $\mathbf{C}^8$ extracted from the trajectory. The diameters of the dots code the sizes of the matrix elements $C^8_{nn'}$. (Matrix elements $C^8_{nn'} < 0.01$ not drawn.)

4. Summary and Conclusion

For the analysis of high-dimensional time series in terms of coarse-grained Markov models we first have applied a density-oriented discretization of the data space. The properties of this partition ensure a balanced statistics for the estimation of all elements of the correspondingly discretized transfer operator $\mathbf{C}$. The nonlinear dynamics eq 8 involving $\mathbf{C}$ was shown to classify the elements $\mathbf{x}_t$ of the time series in terms of prototypical points $\mathbf{y}_\nu^n$ marking the states $n = 1,..., N_\nu$ of coarse-grained Markov models $\mathbf{C}^\nu$. By varying the strength κ of the nonlinearity a hierarchy of such models is obtained, in which the number of states monotonically increases with κ in the range $N_\nu = 1,..., R$. Here, the case $N_\nu = 1$ is the trivial stationary model, and $N_\nu = R$ recovers the original discretization. Two different algorithms have been introduced to construct coarse-grained transfer operators $\mathbf{C}^\nu$ at the intermediate levels ν of the hierarchy. Here, the more time-consuming but accurate approach applies a classification of the data $\mathcal{X}$, whereas the other variant rests on a classification of prototypical points. The correctness of the latter procedure has been demonstrated by comparison with a deterministic and stepwise coarse-graining of the original R-state transfer operator $\mathbf{C} \equiv \mathbf{C}^R$.

For all these approaches observables were introduced, which allow for identifying the most plausible level within the thus constructed hierarchies of models. Their validity has been checked using a synthetic one-dimensional time series, which, apart from its low-dimensionality, exhibits all the characteristics of the relevant model class.

As an example for a more realistic application we have analyzed a six-dimensional time series obtained from a MD simulation of a tripeptide in aqueous solution. In this case the most plausible Markov model could be a priori guessed by physical knowledge on the conformational dynamics of such systems, and our approach actually recovered this guess by analysis of the simulation data. Although in practical applications questions concerning e.g. the number of partition volumes, by which the data space is discretized, the size of the time step, at which a dynamics is sampled, or the validity of the Markovian assumption for the coarse-grained time series have to be additionally addressed, the presented results demonstrate that our approach toward the identification of the most plausible coarse-grained Markov model compatible

with the observations is actually viable. We would like to stress that the approach is applicable also to extremely high-dimensional data sets, which could result, e.g., from simulations of protein folding.

Appendix A: Discrete Dynamics

To solve the differential eq 8 numerically for given $\mathbf{x}_t$ and κ, we use the following algorithm, where $\tau = 1, 2, ...$ denotes discrete time steps of widths $\Delta\tilde{t}$.

- Calculate the probability vector (7).
- While $|\mathbf{p}(\tau + 1) - \mathbf{p}(\tau)| > \epsilon$, $0 \leq \epsilon \ll 1$:

$$\mathbf{p}^{(1)} = \mathbf{C}\mathbf{p}(\tau) \tag{26}$$

$$p_r^{(2)} = p_r^{(1)} + \tilde{\kappa} p_r^{(1)}[p_r^{(1)} - (\mathbf{p}^{(1)})^2] \tag{27}$$

$$p_r^{(3)} = \begin{cases} 0, & \text{if } p_r^{(2)} < 0 \\ 1, & \text{if } p_r^{(2)} > 1 \\ p_r^{(2)}, & \text{otherwise} \end{cases} \tag{28}$$

$$\mathbf{p}(\tau + 1) = \frac{\mathbf{p}^{(3)}}{\sum_{r'} p_{r'}^{(3)}} \tag{29}$$

Equations 26 and 27 discretize the differential eq 8 under the approximation

$$\ln[(\mathbf{C} - 1) + 1] = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}(\mathbf{C}\text{-}1)^n}{n} \approx (\mathbf{C} - 1) \tag{30}$$

and with $\tilde{\kappa} = \Delta\tilde{t}\kappa$. The cutoff (28) and the renormalization (29) serve to stabilize the algorithm numerically.

Appendix B: Attractors of the Nonlinear Dynamics

Here, we consider the nonlinear part

$$\frac{d}{d\tau} p_r = p_r(p_r - \mathbf{p}^2) \tag{31}$$

of (8) for the probability distribution (7). The uniform distribution

$$p_r = \begin{cases} \dfrac{1}{|\mathscr{M}|}, & r \in \mathscr{M} \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

over a nonempty subset $\mathscr{M} \subset \{1,..., R\}$ of indices $r \in \mathscr{M}$ is stationary under (31), as can be seen by inserting (32) into (31). For each $M \equiv |\mathscr{M}|$ there are

$$\binom{R}{M}$$

possibilities to choose an index set $\mathscr{M}$. Therefore, there are a total of

$$\sum_{M=1}^{R} \binom{R}{M} = 2^R - 1$$

stationary solutions of (31). By applying a small deviation to one component p_s, $s \in \mathscr{M}$, one can easily show that only the R δ-distributions ($M = 1$) are stable attractors of (31).

Appendix C: Simulation Method

As a simulation system we have chosen a periodic rhombic dodecahedron (inner radius $R_I = 17$ Å) filled with 930 water molecules and one serine tripeptide molecule with acetylated N- and amidated C-termini. A Berendsen thermostat and barostat[42] were used to control the temperature at 300 K and the pressure at 1 atm. The molecular mechanics of the system was described by means of the all-atom force field CHARMM22.[43] Toroidal boundary conditions were applied to the computation of the electrostatics. As described in detail in refs 44 and 45 they comprise a moving-boundary reaction field description for electrostatic interactions beyond a distance of about R_I and fast hierarchical multipole expansions combined with a multiple time step integrator[46] at smaller distances. The basic integration time step was $\Delta t = 1$ fs. By periodically saving the peptide configuration the sampling time step was set to $\Delta\tilde{t} = 1$ ps.

Acknowledgment. This work was supported by the Deutsche Forschungsgemeinschaft (SFB 533/C1).

Supporting Information Available: Stationary distributions $\mathbf{p}_{\text{stat}}$ (Figure 1), temporal evolution of the virtual density (Figure 2), and relaxation time scales (Figure 3). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hamilton, J. D. *Time series analysis;* Princeton University Press: Princeton, 1994.

(2) Bloomfield, P. *Fourier analysis of time series;* Wiley: New York, 2000.

(3) Percival, D. B.; Walden, A. T. *Wavelet methods for time series analysis;* Cambridge University Press: Cambridge, 2000.

(4) Bradley, E. Analysis of time series. In *Intelligent data analysis*; Berthold, M., Hand, D. J., Eds.; Springer: Berlin, 2003, 199−227.

(5) Rabiner, L. R. *Proc. IEEE* **1989**, *77*, 257−286.

(6) Coast, D. A.; Stern, R. M.; Cano, G. G.; Briller, S. A. *IEEE Trans. Biomed. Eng.* **1990**, *37*, 826−836.

(7) Ephraim, Y.; Merhav, N. *IEEE Trans. Inform. Theory* **2002**, *48*, 1518−1569.

(8) Gardiner, C. W. *Handbook of stochastic methods;* Springer: Berlin, 1990.

(9) Taylor, P. Statistical methods. In *Intelligent data analysis*; Berthold, M., Hand, D. J., Eds.; Springer: Berlin, 2003; pp 69−129.

(10) Branden, C.; Tooze, J. *Introduction to protein structure*; Garland: New York, 1999.

(11) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids;* Clarendon Press: Oxford, 1987.

(12) van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992−1023.

(13) Grubmüller, H.; Tavan, P. *J. Chem. Phys.* **1994**, *101*, 5047−5057.

(14) Giuliani, A.; Manetti, C. *Phys. Rev. E* **1996**, *53*, 6336−6340.

(15) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495−1517.

(16) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.

(17) Huisinga, W.; Best, C.; Roitzsch, R.; Schütte, C.; Cordes, F. *J. Comput. Chem.* **1999**, *20*, 1760−1774.

(18) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear Algebra Appl.* **2000**, *315*, 39−59.

(19) Hamprecht, F. A.; Peter, C.; Daura, X.; Thiel, W.; van Gunsteren, W. F. *J. Chem. Phys.* **2001**, *114*, 2079−2089.

(20) Schütte, C.; Huisinga, W.; Deuflhard, P. Transfer operator approach to conformational dynamics in biomolecular systems. In *Ergodic theory, analysis, and efficient simulation of dynamical systems*; Fiedler, B., Ed.; Springer: Berlin, 2001; pp 191−224.

(21) Carstens, H.; Renner, C.; Milbradt, A.; Moroder, L.; Tavan, P. *Biochemistry* **2005**, *44*, 4829−4840.

(22) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598−1603.

(23) Eckmann, J. P.; Kamphorst, S. O.; Ruelle, D. *Europhys. Lett.* **1987**, *4*, 973−977.

(24) Dersch, D. R.; Tavan, P. Load balanced vector quantization. In *Proc. ICANN'94, Sorrento*; Moreno, M., Morasso, P. G., Eds.; Springer: London, 1994.

(25) Kloppenburg, M.; Tavan, P. *Phys. Rev. E* **1997**, *55*, 2089−2092.

(26) Albrecht, S.; Busch, J.; Kloppenburg, M.; Metze, F.; Tavan, P. *Neural Networks* **2000**, *13*, 1075−1093.

(27) Hillermeier, C.; Kunstmann, N.; Rabus, B.; Tavan, P. *Biol. Cybern.* **1994**, *72*, 103−117.

(28) Kohonen, T. *Biol. Cybern.* **1982**, *43*, 59−69.

(29) Kohonen, T. *Biol. Cybern.* **1982**, *44*, 135−140.

(30) Dersch, D. R.; Tavan, P. *IEEE Trans. Neural Networks* **1995**, *6*, 230−236.

(31) Haken, H. *Synergetics;* Springer: Berlin, 1983.

(32) Dellnitz, M.; Hohmann, A.; Junge, O.; Rumpf, M. *CHAOS: Interdiscip. J. Nonlinear Sci.* **1997**, *7*, 221−228.

(33) Duda, R. O.; Hart, P. E. *Pattern classification and scene analysis;* Wiley: New York, 1973.

(34) The optimization of $\tilde{p}(\mathbf{x}|\mathcal{W};\sigma)$ can be interpreted as a self-organizing learning process of forward connections within a two-layer generalized radial basis functions (GRBF) network.[26,47]

(35) Voronoi, G. F. *J. Reine Angew. Math.* **1908**, *134*, 198−287.

(36) The matrix **C** can be interpreted as a self-organizing matrix of lateral connections in the central layer of a recurrent three-layer GRBF network (cf. ref 27).

(37) In the following the parameter τ denotes the time of a probability dynamics. It should not be confused with the time steps t of the analyzed time series.

(38) In the theory of neural networks the centers $\mathbf{w}_r$ of the Gaussians in (2) are also called the *virtual positions* of the associated mapping neurons.[48]

(39) The $\mathbf{y}_\kappa^n$ can be interpreted as the output of the neural network addressed in ref 36.

(40) Jammalamadaka, S. R.; Gupta, A. S. *Topics in circular statistics;* World Scientific: Singapore, 2001.

(41) Hu, H.; Elstner, M.; Hermans, J. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 451−463.

(42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(43) MacKerell, A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(44) Mathias, G.; Tavan, P. *J. Chem. Phys.* **2004**, *120*, 4393−4403.

(45) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003**, *118*, 10847−10860.

(46) Eichinger, M.; Grubmüller, H.; Heller, H.; Tavan, P. *J. Comput. Chem.* **1997**, *18*, 1729−1749.

(47) Bishop, C. *Neural networks for pattern recognition*; Clarendon Press: Oxford, 1997.

(48) Tavan, P.; Grubmüller, H.; Kühnel, H. *Biol. Cybern.* **1990**, *64*, 95−105.

CT050020X

Supplementary Information for:

Extracting Markov models of peptide conformational dynamics from simulation data

Verena Schultheis, Thomas Hirschberger, Heiko Carstens, Paul Tavan

Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität,
Oettingenstr. 67, 80538 München, Germany

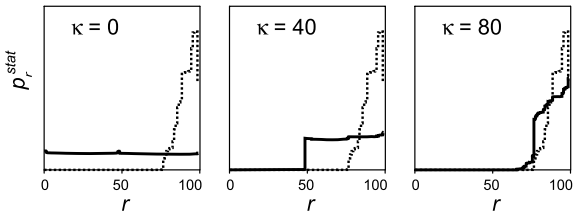

Figure 1: Solid lines: Stationary distributions $\mathbf{p}_{stat}$ for $\kappa \in \{0, 40, 80\}$ and $x_t = 0.66$ as in Fig. 4 of the manuscript. At $\kappa = 0$ the peaked initial distribution (dashed lines) converges to the uniform distribution, at $\kappa \in \{40, 80\}$ it becomes more and more localized near the intermediate metastable states of the linear dynamics (cf. Fig. 4 of our article) in the right half of the 100-dimensional state space.

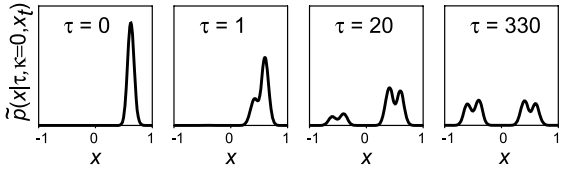

Figure 2: Temporal evolution of the virtual density $\tilde{p}(\mathbf{x}|\tau, \kappa = 0, x_t)$ for a linear dynamics (cf. Fig. 4 of the manuscript). As dictated by the starting value $x_t = 0.66$ the density is initially ($\tau = 0$) concentrated near the right maximum $x_4 = 0.6$ of the data distribution (cf. Fig. 2 of our article) and immediately starts spreading towards the mixture model (2) of the invariant density $p_{inv}(\mathbf{x})$.

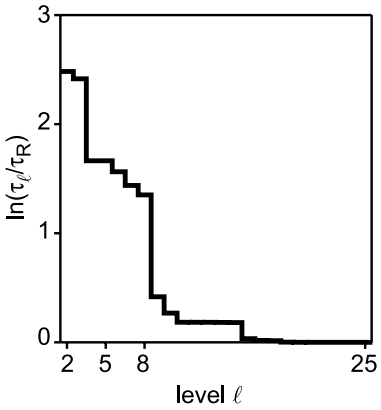

Figure 3: Logarithmic plot of the fastest relaxation time scales τ_ℓ measured in units of τ_R at the 24 steps of the recursive coarse-graining procedure applied to the transfer operator of the tripeptide. The figure is analogous to Fig. 9 of our article, which pertains to the one-dimensional model trajectory. In the transition from hierarchy level $\ell = 9$ to $\ell = 8$ a large increase of the fastest relaxation time scale is observed, which indicates that an eight-state model should describe the key features of the conformational dynamics in the peptide.

3 Ein polarisierbares Kraftfeld zur Berechnung von Infrarotspektren des Proteinrückgrats

Dieses Kapitel beschreibt ein Kraftfeld für das Proteinrückgrat, dessen Parameter wie Bindungslängen und -winkel vom anliegenden elektrischen Feld abhängen. Es ist ein Abdruck des Manuskripts [1]

> Verena Schultheis, Rudolf Reichold, Bernhard Schropp, and Paul Tavan: „A polarizable force field for computing the infra-red spectra of the polypeptide backbone.“,

das ich gemeinsam mit Rudolf Reichold, Bernhard Schropp und Paul Tavan zur Veröffentlichung im *Journal of Physical Chemistry B* eingereicht habe. Zu dieser Veröffentlichung gehört Zusatzmaterial in Form ergänzender Grafiken, die im Anschluss an das Manuskript ebenfalls abgedruckt sind.

[1] Reproduced with permission from the the Journal of Physical Chemistry B, in press. Unpublished work copyright 2008 American Chemical Society.

A polarizable force field for computing the infra-red spectra of the polypeptide backbone

Verena Schultheis, Rudolf Reichold, Bernhard Schropp, and Paul Tavan*
Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität,
Oettingenstr. 67, 80538 München, Germany

*email: tavan@physik.uni-muenchen.de; phone: +49-89-2180-9220, fax: +49-89-2180-9202

Abstract

The shapes of the amide bands in the infrared (IR) spectra of proteins and peptides are caused by electrostatically coupled vibrations within the polypeptide backbone and code the structures of these biopolymers. A structural decoding of the amide bands has to resort to simplified models, because the huge size of these macromolecules prevents the application of accurate quantum mechanical methods such as density functional theory (DFT). Previous models employed transition-dipole coupling methods that are of limited accuracy. Here we propose a concept for the computation of protein IR spectra, which describes the molecular mechanics (MM) of polypeptide backbones by a polarizable force field of "type II". By extending the concepts of conventional polarizable MM force fields such a PMM/II approach employs field dependent parameters not only for the electrostatic signatures of the molecular components but also for the local potentials modeling the stiffness of chemical bonds with respect to elongations, angle deformations and torsions. Using a PMM/II force field the IR spectra of the polypeptide backbone can be efficiently calculated from the time dependence of the backbone's dipole moment during a short (e.g. 100 ps) MD simulation by Fourier transformation. PMM/II parameters are derived for harmonic bonding potentials of amide groups in polypeptides from a series of DFT calculations on the model molecule N-methyl acetamide (NMA) exposed to homogeneous external electric fields. The amide force constants are shown to vary by as much as 20 % for relevant field strengths. As a proof of principle it is shown that the large solvatochromic effects observed in the IR spectra of NMA upon transfer from the gas phase into aqueous solution are not only excellently reproduced by DFT/MM simulations but are also nicely modeled by the PMM/II approach. The tasks remaining for a proof of practice are specified.

Keywords: amide bands, molecular dynamics simulation, DFT/MM hybrid methods, normal mode analyses, N-methyl acetamide

1 Introduction

The infra-red (IR) spectra of proteins and peptides are dominated by the so-called amide bands. These bands are due to vibrations of the strongly polar and polarizable amide groups (AGs) C_α–C′O–NH–C_α making up the polypeptide backbone. The most prominent representative is the so-called amide I (AI) band, which belongs to the C′=O stretching motions. Because of dipole-dipole coupling the corresponding normal modes are delocalized over several AGs. This coupling is steered by the relative orientations and distances of the AGs. Therefore, the spectral positions and shapes of the amide bands code the three-dimensional structures of the polypeptides.[1–3]

As we will discuss in more detail further below, the derivation of accurate structural information from observed amide bands poses a major challenge. Up to now the structural decoding of amide spectra still has to rely on empirical rules derived from correlations between observed amide band shapes and structural information obtained by other means (e.g. by x-ray or multi-dimensional NMR).[1,2] These rules allow us to roughly estimate a polypeptide's

content in secondary structure motifs – such as α-helices, β-sheets or loops – from the IR spectrum. However, for quantitative analyses of such spectra a well-founded and reliable theoretical method is needed. It is the aim of this paper to pave the way towards the construction of a corresponding computational approach.

This article is organized as follows: We first analyze the problem and the state of the art by discussing well-known experimental data and previous theoretical concepts. This introductory review serves to explain our working hypothesis that a new type of polarizable force field (called PMM/II) should be constructed for AGs to enable the computation of peptide IR spectra from molecular dynamics (MD) simulations. Subsequently we will outline the basic considerations that serve as guidelines in our design of a PMM/II force field for AGs. After a general sketch of how the parameters of such a force field can be derived from quantum mechanical calculations applying density functional theory, we give details of the employed computational methods. Our presentation of results starts with a check of how well our particular DFT approach performs on the vibrational spectrum of an isolated NMA molecule. Having thus established a quantum mechanical reference we check the quality at which our most simple PMM/II force field describes the vibrational spectrum of an AG in the zero-field case and discuss future extensions. Next we present the response properties of the force field parameters to external fields and study how well a DFT/MM description of the IR spectrum of protonated and deuterated NMA in D_2O is reproduced by our current PMM/II approach. A short outline of the tasks remaining for a proof of practice concludes the paper.

2 The nature of the problem

In principle the vibrational spectra of AGs should be calculated by quantum mechanical methods. Here, density functional theory (DFT)[4,5] currently offers the best trade-off between accuracy and computational efficiency.[6–8] By combining the DFT treatment of a molecule with a parameterized molecular mechanics (MM) model of its condensed phase environment one can account for the polarization of the given molecule by this environment.[8,9] In fact, for model peptides the use of DFT methods has yielded important insights into the amide spectra (see e.g. Ref. 10). Furthermore, a DFT/MM approach has been successfully applied to derive the ongoing structural changes from time-resolved IR spectra for the sub-nanosecond unfolding of a light-switchable β-hairpin peptide.[11] But even for the very small peptides considered in the quoted studies the computational effort is tremendous and becomes rapidly unfeasible for larger peptides or for proteins.

This is why simplified procedures have been sought that can avoid the intractable computational effort associated with quantum mechanical descriptions of the polypeptide IR spectra. Early suggestions for such models combined the large transition dipole moments of the local C′=O stretching modes through dipole-dipole coupling into normal modes widely delocalized over the backbone. These so-called *transition dipole coupling* (TDC) models assumed that all frequencies of the local C′=O stretching motions are identical (or that all hydrogen-bonded C′=O dipoles vibrate with one frequency, and all solvent-exposed C′=O dipoles with another).[12–14] Thus, the TDC models tried to explain the shapes of the observed AI bands mainly through the distance- and orientation-dependent couplings of the local C′=O dipoles.

2.1 Effects of Polarization in AGs

However it turned out that these TDC approaches were incapable to completely account for the observed band shapes.[8,12–19] Therefore, starting in 2002, the TDC approach was extended[15–19] by explicitly including field-induced C′=O frequency changes into the description. These extended methods, which go back to the work of Cho et al.[15,16] and have been taken up by others,[17–19] thus abandoned the simplifying assumption of identical C′=O frequencies.

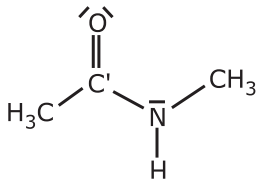

Figure 1: The molecule N-methyl acetamide (NMA), which is a model compound for the amide groups (AGs) in proteins.

At this point we would like to stress that there was never any good reason to consider the assumption of identical C′=O frequencies as plausible. As

we will explain, this judgment immediately follows from the well-known vibrational spectra[20–23] of the most simple molecular model of an AG, which is the molecule N-methyl acetamide (NMA) depicted in Fig. 1.

In the gas phase the AI band of NMA lies at around 1730 cm^{-1}, in n-hexane at 1697 cm^{-1}, in acetonitrile at 1674 cm^{-1}, in dimethylsulfoxide at 1667 cm^{-1}, and in water at 1625 cm^{-1}.[20–23] According to these data, the solvent-induced red-shifts of the AI band increase with the polarity of the solvent and span up to about 100 cm^{-1}. Clearly, these red-shifts must be due to the electrostatic reaction field that is generated by the solvent dipoles (oriented in response to NMA's strong dipole moment) and acts on the polarizable NMA molecule. Therefore, the strong solvatochromism of the AI band indicates that the force constant of the C′=O stretching mode must be highly sensitive to the local electric field acting on NMA.

Because the structures of polypeptides are mainly shaped by complex electrostatic interactions within these biological macromolecules and with the surrounding solvent, one expects sizable variations among the electric fields acting on its various AGs.[8] Taking NMA as a model for these AGs one thus expects that the frequencies of the local C′=O oscillators along the backbone of a polypeptide should exhibit considerable field-induced differences, in contrast to the basic assumption of the early TDC models.

Interestingly, the amide II (AII) and amide III (AIII) bands of NMA, which mainly involve the C′–N stretching and N–H in-plane bending motions,[23] show an inverse solvatochromism: Here, the least polar medium, i.e. the gas phase, is associated with low frequencies of 1500 cm^{-1} (AII) and 1257 cm^{-1} (AIII), the mildly polar solvent acetonitrile leads to higher frequencies of 1546 cm^{-1} (AII) and 1285 cm^{-1} (AIII), and in the strongly polar solvent H_2O the frequencies are further up-shifted to 1582 cm^{-1} (AII) and 1317 cm^{-1} (AIII).[20–23] Thus, in polar media these bands shift to the blue by up to 82 cm^{-1} (AII) and 60 cm^{-1} (AIII).

The large solvatochromic effects observed for NMA can be qualitatively explained by the resonance structures of the π electron system, which this molecule shares with every AG. These resonance structures are depicted in Fig. 2. Due to the π resonance, the molecular wave function is a coherent superposition of a neutral (A) and a zwitterionic (B) structure. While structure A has a double bond between the central carbon atom C′ and the oxygen atom O, structure B exhibits a double bond between C′ and the nitrogen atom N. The resulting partial double bonds cause the amide atoms to form a quite rigid plane (indicated by dashed lines), which we can describe by the Cartesian coordinate system specified in the figure. The y-axis is parallel to the C′=O bond, the x-axis is perpendicular to this bond in the O=C′–N plane, and the z-axis (not drawn) points into this plane. In structure B the oxygen atom O carries a negative and the hydrogen atom H a positive charge. The coherent admixture of structure B to the wave function thus additionally explains the strong electric dipole moments of the AGs.

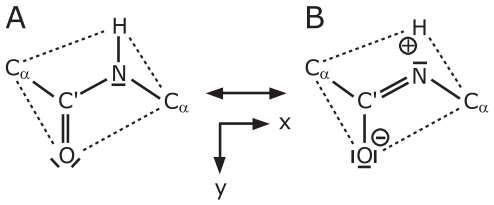

Figure 2: The two π electron resonance structures (A: neutral, B: zwitterionic) of an AG as found in the backbone of a polypeptide (drawing adopted from Ref. 8). Upon replacing the C_α atoms by methyl groups one obtains the model compound NMA (cf. Fig. 1). A Cartesian coordinate system is attached to the AG as indicated.

If an external electric field $E_y < 0$ oriented antiparallel to the y-axis acts on an AG, the contribution of resonance structure B to the wave function will increase because the associated electric dipole is aligned with the external field in an energetically favorable way. The increased admixture of B has two consequences, which can be immediately identified by looking at Fig. 2: Because in B the C′–O bond is a single bond, the equilibrium bond length $l_{C'O}$ will increase and the bond will become less stiff, i.e. its force constant $k_{C'O}$ will decrease. For the C′–N bond we expect exactly the opposite effects.

The discussed direction $(-y)$ of the electric field is the generic case for the reaction fields that an electric dipole – such as the dipole of the AG – causes in polar media. Therefore the resonance structures in Fig. 2 can actually explain why in polar media the AI band of NMA, which is associated with the C′=O stretching motion, shifts to the red, while the AII and AIII bands, which both involve the C′–N stretching motion, shift towards the blue. Moreover the simple resonance considerations have addition-

ally shown that the whole intramolecular force-field of an AG (and not only the C′=O stretching force constant) is expected to vary with an external electric field. Quite obviously these strong polarization effects should not be neglected in computations of amide spectra – in contrast to the assumptions of the early TDC models.

2.2 Polarizable TDC models and their limitations

However, also the mentioned extensions of the TDC approach,[15–19] which aim at including polarization effects, do not completely live up to the insights that emerged from the above discussion of the resonance structures. These extensions restrict the field dependence (as determined by quantum chemical calculations on NMA-water clusters, NMA-methanol clusters, di-, tri-, and polypeptides[15–19]) to the C′=O frequencies – instead of including the field dependence of the *whole* amide force field (for an overview about the series of related publications we refer to Refs. 16 and 18).

Within the framework of these extended approaches (and still following the basic TDC concept) one sets up for any given structure of a polypeptide an excitonic coupling matrix, whose diagonalization yields the AI combination modes.[15–19] If one repeats this procedure for a series of structural snapshots from a molecular dynamics (MD) simulation, one obtains — just like in an "instantaneous normal mode analysis" — an AI spectrum, which is inhomogeneously broadened as a result of structural fluctuations. The resulting band widths, however, will be largely overestimated, because the effects of motional narrowing are neglected.[8,17,30,31]

Despite the latter criticism the quoted extensions of the original TDC approach definitely represent a considerable progress. Correspondingly the calculated AI bands agree much better with experimental data.[15–19] On the other hand already the wide range of different parameterizations documented in the various papers[16,17] suggests that also the extended TDC approach does not yet adequately cover the underlying physics.

Concerning this matter the detailed quantum mechanical study on NMA by Kubelka and Keiderling[20] contains an important hint. This study points to the significance of the spectral distance between the AI and AII modes for correctly calculating the AI frequency. As explained by the authors, the AI frequency can be accurately determined only if the AI-AII spectral distance is correctly described. However, as we have seen above in our discussions of the experimental data on NMA and of the resonance structures, the electronic polarization affects the AI and AII bands contrarily. Therefore, it seems to be inevitable that one has to explicitly include the field dependence of the AII mode and of its coupling to the AI mode instead of solely trying to parameterize the field dependence of the AI frequency. This line of argumentation can be extended to the AII and AIII modes, and so forth. As a result one comes to the conclusion that one should account for the field dependence of the whole force field of the AGs even if, in the case of NMA, one wants to understand only the solvatochromism of the AI band or, in the case of polypeptides, only the complex shape and spectral position of this particular band.

2.3 The resulting task

The above conclusion formulates the task, which we want to tackle in this work. It is the task of constructing a new type of polarizable force field for AGs, in which all parameters that serve to model local chemical bonding forces are chosen field-dependent. Conventional polarizable MM force fields (PMM), in contrast, restrict the field-dependence to the electrostatic signatures of molecular structures (e.g. through complementing the usual set of fixed atomic partial charges by inducible point-dipoles).[24–29] To stress this important difference, we will call conventional PMM force fields as "type I" and the extended ones as "type II".

Such a PMM/II force field can become a viable approach, only if the chemical bonding forces show a linear response to the locally acting electric field, because only then computationally efficient implementations suited for large scale MD simulations are readily constructed. PMM/I force fields generally rely on linear response for the computation of induced dipole moments or of charge fluctuations,[24–29] and it is a first task to check whether linear response also applies to changes of bonding properties, if the external fields have the typical strengths, which occur in condensed phase environments.

As soon as such a PMM/II force field for the peptide backbone will exist, one can employ it in MD simulations of a protein in solution. Then one can calculate the IR spectrum by Fourier transform or generalized viral frequency techniques from the MD trajectories thereby including the effects of mo-

tional narrowing (see Ref. 31 for details). Therefore, our approach concurrently implies that we will abandon the excitonic TDC concept, which was inherently restricted to the AI band. Through the use of MD simulation techniques, which automatically account for all electrostatic interactions within a protein and, thus, also for the dipolar couplings among the AG's, the complete vibrational spectrum of the polypeptide backbone will become accessible instead.

3 Guidelines and Procedures of Design

Any construction of a MM force field has to start with a set of choices concerning the complexity of the model and the methods for computing its parameters. Preferentially such a model should be as simple as possible, because simplicity is closely related with computational efficiency. Furthermore, its ingredients (i.e. the various employed model potentials) should have clear physical correlates, because the existence of such correlates can simplify the computation and optimization of the parameters. On the other hand, the model should be complex enough to enable sufficiently accurate descriptions of the quantities of interest.

In the given case the quantities of interest are the vibrational spectra of complex molecular structures, i.e. of the backbones of proteins and peptides. To enable interpretations of experimental data, the desired computational approach must provide access to fine details of these spectra. Because these details quite sensitively depend on the frequencies and couplings of the local oscillators, the given task of constructing a "sufficiently accurate" force field becomes a particularly demanding challenge.

3.1 Simplifying assumptions and decomposition of the task.

Despite the obvious intricacy of our task, there are a several aspects nourishing the hope that it can be simplified.

1. Polypeptide backbones are polymers composed of identical units, i.e. of the six-atomic structural core motifs C_α–C′O–NH–C_α defining the AGs (cf. Fig. 2). If one assumes that the couplings between the various AGs within a polypeptide backbone are dominated by electrostatic interactions, then one may hope that an accurate PMM/II model for any given AG, which properly accounts for the AG's response to the locally acting electric field and for its electrostatic signatures, can enable the description of the backbone IR spectra through application of Fourier transform techniques to PMM/II-MD trajectories.

2. With the molecule NMA depicted in Fig. 1 is a most simple model for the AG building block of polypeptides available and the vibrational spectrum of this model is experimentally quite well characterized.[20–23] However, because the vibrational modes localized within the methyl groups of NMA can mix with the twelve normal modes of the AG core motif, the corresponding "amide modes" are not easily identified within the NMA spectrum. For the purpose of such identification one must provide procedures, by which the IR spectra of NMA can be reliably calculated and, subsequently, can be reduced to those of its AG core.

3. As mentioned in Sec. 2, the gas phase IR spectra of organic molecules like NMA can be speedily and accurately computed by DFT methods.[6–8] Due to a fortuitous cancellation of errors, specifically the BP86 functional[32,33] is known to yield harmonic intra-molecular force fields matching the frequencies of the observed anharmonic fundamentals particularly well.[7] Thus, BP86 seems to be the functional of choice, if one needs a quantum mechanical reference for the MM construction of a largely harmonic AG force field that is supposed to serve for the computation of vibrational spectra. This quantum mechanical reference is, of course, the Hessian matrix $\mathbf{H}$ calculated by BP86 for an NMA molecule at its equilibrium geometry $\mathcal{G}$. If one suitably restricts $\mathbf{H}$ to the degrees of freedom characterizing the AG core, one can gain access to the harmonic force field of a single AG.

4. Furthermore, also the response of the AG force field to external electric fields $\mathbf{E} \neq 0$ can be calculated by DFT. As long as that response is linear, it can be uniquely characterized by exposing NMA to homogeneous electric fields of varying strengths oriented parallel to the axes of the coordinate system depicted in Fig. 2. Here, the homogeneity of the applied electric fields is important, because only then one can gain insights into the proper treatment of field

inhomogeneities. In condensed matter such inhomogeneities are inevitable. They cause the problem that the strength of the field polarizing a molecule, which is a volume average, may strongly and systematically deviate from the field strength encountered at a specific position within that molecule.[34] In the case of liquid water, e.g., the external field at the oxygen atom is on average by about 40 % larger than the volume average field which actually polarizes a given water molecule in the aqueous environment.[34] When constructing PMM force fields, in which, for computational reasons, the strengths of polarizing fields cannot be calculated through volume integrals, this issue must be carefully dealt with.

The above arguments suggest that the determination of a PMM/II force field for polypeptide backbones can be reduced to an extended series of DFT calculations on NMA. These computations should yield the equilibrium geometries $\mathcal{G}(\mathbf{E})$ and Hessian matrices $\mathbf{H}(\mathbf{E})$ of NMA exposed to constant electric fields $\mathbf{E}$ of varying strengths and directions. Here, the case $\mathbf{E} = 0$ marks the gas phase limit, in which NMA is characterized by $\mathcal{G}^0$ and $\mathbf{H}^0$. The data on the NMA properties $\mathcal{G}(\mathbf{E})$ and $\mathbf{H}(\mathbf{E})$ then represent the basis for the computation of a PMM/II force field for the AGs of polypeptides.

3.2 Transforming DFT data into an AG force field.

Standard MM force fields such as CHARMM22[35] employ simple harmonic model potentials for bond stretches and bond angle deformations. They avoid cross-terms, by which one can include dynamical couplings between these internal coordinates. Although such cross-terms are vital for an accurate modeling of a DFT Hessian,[25,36,37] we will restrict this first attempt of constructing a PMM/II force field to CHARMM-type potentials, because then it can be much more rapidly and safely implemented into our parallelized MD-program package EGO.[38]

The restriction to harmonic model potentials for a non-redundant set of internal coordinates has the additional advantage that the parameterization of a PMM/II force field for AGs from the DFT references $\mathbf{H}(\mathbf{E})$ becomes very simple. Then one can use a method suggested by Boatz and Gordon[39] and derive from the Hessians $\mathbf{H}(\mathbf{E})$ for each internal coordinate q_i a so-called "intrinsic" frequency $\nu_i^q[\mathbf{H}(\mathbf{E})]$, which uniquely determines the parameters of the harmonic model potential associated with this internal coordinate. We will now shortly sketch a non-redundant set of internal coordinates for an AG, the associated force field and a procedure for parameter computation based on intrinsic frequencies.

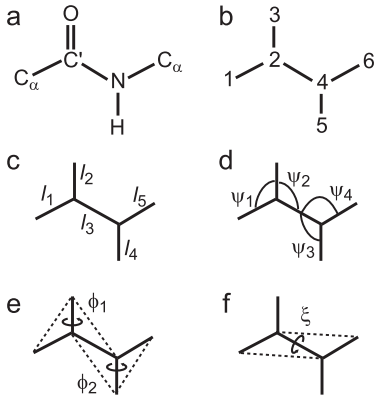

Figure 3: A non-redundant set of twelve internal coordinates for a simple AG force field. a) Atoms and b) their numbering. The AG geometry is uniquely given by c) five bond lengths l_i, d) four bond angles ψ_i, e) two improper dihedral angles ϕ_i, and f) one proper dihedral angle ξ.

3.2.1 A non-redundant set of internal coordinates.

To cover the twelve internal degrees of freedom q_j, $j = 1, \ldots, 12$, belonging to the six atoms of an AG we have chosen the internal coordinates depicted in Fig. 3. Here, the identification of the first five internal coordinates with the five bond lengths ($q_j \equiv l_j$, $j = 1, \ldots, 5$) is uniquely fixed by the structure, whereas the choice of the four bond angles $q_{5+i} \equiv \psi_i$, $i = 1, \ldots, 4$, from the six existing candidates is a matter of taste. Our selection of the three additionally required dihedral angles follows from the physical consideration that an AG consists of two local planes defined by the sp^2-hybridizations of the central atoms C′ and N, and that a *cis* to *trans* isomerization involves a rotation of these planes around the C′—N bond by 180°. Correspondingly, the dihedral angles $q_{9+i} \equiv \phi_i$, $i = 1, 2$, measure, to what extent the sp^2-hybridizations of the C′ and N atoms actually enforce locally planar bonding patterns (with planarity given by $\phi_i = 0$). Such dihedral angles are commonly called "improper" to distinguish them from the "proper" dihedral angles

describing actual rotations around chemical bonds. The coordinate $q_{12} \equiv \xi$ is such a proper dihedral and characterizes *cis*- and *trans*-isomers of an AG through the values $\xi = 0$ and $\xi = 180°$, respectively.

3.2.2 A largely harmonic force field.

The set $\mathbf{q}$ of the internal coordinates q_j characterized by Fig. 3 can be employed to set up a simple energy function

$$E_b(\mathbf{q}|\Theta) = E_d(\xi|\theta_\xi) + \sum_{j=1}^{11} E_h(q_j|\theta_j) \qquad (1)$$

for the chemical bonding forces, which is specified by the parameter set $\Theta = \theta_\xi \cup \{\theta_j | j = 1, \ldots, 11\}$. This function is composed of only one non-harmonic contribution

$$E_d(\xi|\theta_\xi) = \frac{E_1}{2}\left[1 + \cos(\xi)\right] + \frac{E_2}{2}\left[1 + \cos(2\xi + 180°)\right] \qquad (2)$$

for the single torsional degree of freedom within the system and of eleven harmonic potentials

$$E_h(q_j|\theta_j) = \frac{k_j}{2}(q_j - q_j^0)^2, \quad j = 1, \ldots, 11, \qquad (3)$$

for the remaining internal coordinates q_j.

The non-harmonic contribution (2) is specified by the parameters $\theta_\xi = \{E_1, E_2\}$. For E_1, $E_2 > 0$ it is a model potential that distinguishes the *cis*- and *trans*-isomers of an AG as two stable configurations. The energy $E_d(\xi = 0) = E_1$ of the *cis*-state is assumed to be above the energy $E_d(\xi = 180°) = 0$ of the *trans*-state, and the two states are assumed to be separated by a rotational energy barrier. Note, however, that in the vicinity of the *trans*-configuration ($\xi \approx 180°$) the model potential (2) approximately reduces to a harmonic potential

$$E_d(\xi|\theta_\xi) \approx \frac{k_\xi}{2}(\xi - 180°)^2, \qquad (4)$$

whose force constant $k_\xi = (E_1/2 + 2E_2)/\mathrm{deg}^2$ is determined by the parameters θ_ξ. If one knows the *cis*-*trans* energy difference E_1 (e.g. from a DFT calculation) and if one is mainly interested in a model potential, which is accurate at small deviations $|\xi - 180°|$, the missing parameter E_2 can be determined by tuning the harmonic force constant k_ξ in Eq. (4) by the same methods, by which one also tunes the force constants k_j of the other harmonic potentials.

Thus Eqs. (4) and (3) associate to every internal coordinate q_j of a *trans*-AG a harmonic potential characterized by parameters $\theta_j = \{k_j, q_j^0\}$. For most of these internal coordinates the harmonic approximation should be reasonable, because at physiological temperatures a thermal excitation will not cause large deviations of the q_j from their equilibrium values q_j^0. For the bond lengths l_i and angles ψ_i this expectation of a stiff binding to the equilibrium values l_i^0 and ψ_i^0 is well-founded by a large body of experience with small molecules like NMA. For the potentials associated with the two improper dihedral angles $q_{10} \equiv \phi_1$ and $q_{11} \equiv \phi_2$, which measure the stiffness of the local planes defined by the sp^2-hybridizations of the C′ and N atoms, this expectation remains to be checked by computations of the associated force constants k_{10} and k_{11}. Similar considerations apply to the potential (4) for ξ.

On the other hand and as mentioned above, the simple ansatz (3) represents particularly for the ICs q_j, $j = 1, \ldots, 9$, associated with the bond lengths and bond angles an oversimplification, because it neglects the mutual dynamical couplings

$$E_c = \sum_{i<j} k_{ij}(q_i - q_i^0)(q_j - q_j^0) \qquad (5)$$

whose inclusion is of key importance for an accurate MM modeling of vibrational frequencies.[25,36,37] In the given case of an AG, the inclusion of the couplings (5) requires the estimate of 36 additional coupling force constants k_{ij} and of their field dependence from the DFT Hessians $\mathbf{H}(\mathbf{E})$. Because the aim of this work is mainly to obtain (i) a first estimate on the order of magnitude by which external fields typically occurring in the condensed phase can change the force constants in an AG, (ii) a first answer to the question whether these changes obey linear response, and (iii) a first check on how this polarization effect can be described within a MD simulation approach, we currently avoided the additional complexities of implementing the energy terms (5) and of estimating the additional parameters $k_{ij}(\mathbf{E})$.

3.2.3 A procedure for parameter estimation.

Once one has calculated by DFT the geometry $\mathcal{G}(\mathbf{E})$, the Hessian $\mathbf{H}(\mathbf{E})$, and the intrinsic frequencies $\nu_i^q[\mathbf{H}(\mathbf{E})]$ of the internal coordinates q_i for an NMA molecule in an external field $\mathbf{E}$, the parameters $\theta_j(\mathbf{E}) = \{k_j(\mathbf{E}), q_j^0(\mathbf{E})\}$ of the harmonic potentials (3) and (4) are readily determined by the following procedure:

1. The equilibrium bond lengths $l_i^0(\mathbf{E})$ and bond angles $\psi_i^0(\mathbf{E})$ are taken from $\mathcal{G}(\mathbf{E})$; for the har-

monic potentials involving dihedral angles the equilibrium values q_j^0 are determined by requiring the AG to be planar .

2. A Hessian $\mathbf{H}_{\mathrm{MM}}(\mathbf{E})$ is calculated with a trial set of force constants $k_j(\mathbf{E})$ from the MM force field $E_b[\mathbf{q}|\Theta(\mathbf{E})]$.

3. For each internal coordinate q_j the intrinsic frequency $\nu_j^q\,[\mathbf{H}_{\mathrm{MM}}(\mathbf{E})]$ is repeatedly calculated from $\mathbf{H}_{\mathrm{MM}}(\mathbf{E})$, while the force constant $k_j(\mathbf{E})$ is individually varied until the deviation of $\nu_j^q\,[\mathbf{H}_{\mathrm{MM}}(\mathbf{E})]$ from the DFT target value $\nu_j^q\,[\mathbf{H}(\mathbf{E})]$ falls below a predefined threshold.

Because this simple procedure directly transforms the results of a DFT calculation on NMA into a parameter set Θ of our AG force field, it is well suited to determine the field dependence $\Theta(\mathbf{E})$ of the parameter set. The normal modes $\mathbf{n}_i$ and associated vibrational frequencies ν_i^n of an isolated AG are accessible from the gas phase DFT results $\mathcal{G}^0$ and $\mathbf{H}^0$. To what accuracy these values can be reproduced using the simple harmonic model with the zero-field parameters Θ^0 remains to be seen.

4 Methods

As outlined above, the calculation of the PMM/II parameters $\Theta(\mathbf{E})$ requires the DFT computation of the equilibrium geometries $\mathcal{G}(\mathbf{E})$ and Hessians $\mathbf{H}(\mathbf{E})$ of NMA exposed to homogeneous electric fields. There are (at least) two DFT programs, which provide a simple access to such computations, one being TURBOMOLE[40,41] and the other CPMD[42] in combination[9] with the MM-MD package EGO.[38] Because the latter combination additionally offers the possibility of DFT/MM simulations, by which one can check the performance of the PMM/II force field, we mainly used EGO/CPMD.

4.1 DFT calculations on NMA

To calculate $\mathcal{G}(\mathbf{E})$ and $\mathbf{H}(\mathbf{E})$ for NMA by EGO/CPMD we employed Becke's gradient-corrected exchange functional,[32] Perdew's correlation functional,[33] and the norm-conserving pseudo-potentials of Martins and Troullier.[43] NMA was placed into a rectangular box with the dimensions $11.0 \times 10.5 \times 8.0$ Å^3, such that no atom came closer than 3Å to the surface of the box. Within the box the Kohn-Sham orbitals were expanded into a plane-wave basis set, whose size is given by the cutoff of 80 Ry. We denote this particular DFT approach, whose accuracy for the DFT/MM computation of IR spectra has been demonstrated in several earlier applications,[11,44,45] as "MT/BP". To verify the quality of the MT/BP descriptions delivered by EGO/CPMD for the case of NMA we additionally applied the all-electron program TURBOMOLE with the Gaussian basis set TZVP[46] and the standard BP86 functional.[32,33] Furthermore we employed this BP86/TM approach to compute a rotational potential curve $E_d(\xi)$ connecting the *cis*- and *trans*-isomers of NMA, because EGO/CPMD is not very comfortable for this purpose.

For the various required parameters (e.g. convergence criteria, step sizes of numerical differentiation, etc.) we generally used the default values suggested by the program packages. For TURBOMOLE these defaults are documented in the manual,[40] for EGO/CPMD in Ref. 11. How one can easily import homogeneous electric fields into a DFT Hamiltonian when using EGO/CPMD is elaborated in Ref. 34. Note that in EGO/CPMD the orientation of a molecule exposed to an external electric field can be maintained during geometry optimizations through a rotational (and translational) correction.

4.2 MM models for NMA

To enable comparisons of DFT with MM descriptions of the NMA vibrational spectra we had to set up an MM model for an isolated NMA molecule. For the AG of NMA we chose the force field given by Eq. (1) and for the methyl groups a CHARMM22[35] description that was simplified by omitting suggested Urey-Bradley terms.[47] The harmonic zero-field parameters $\Theta^0 = \{q_j^0, k_j | j = 1, \ldots, 12\}$ of the AG force field were determined from MT/BP results on the equilibrium geometry $\mathcal{G}^0$ and Hessian $\mathbf{H}^0$ of an isolated *trans* NMA molecule using the procedures described in Sec. 3.2. The *cis-trans* energy difference E_1, which is additionally required to specify the dihedral potential Eq. (2), was taken from a BP86/TM *cis-trans* potential curve.

To cover also van der Waals and electrostatic interactions we partially took over the CHARMM22 description. Thus, we adopted the corresponding Lennard-Jones potentials for all six amide atoms (cf. Fig. 2) and omitted such potentials for the methyl hydrogens. Similarly we chose vanishing partial charges for the methyl hydrogens and for the remaining atoms of NMA the values listed in the first data column Q_{pol} of Tab. 1. These values were derived from the DFT/MM simulation of NMA in

Table 1: Partial charges Q_i for NMA.

atom	$Q_{\rm pol}$	$Q_{\rm CHARMM}$
C(C′)	0.000	0.000
O	−0.781	−0.510
C′	0.781	0.510
N	−0.588	−0.470
H	0.354	0.310
C(N)	0.234	0.160

aqueous solution sketched further below by taking averages of ESP charges[9,48] and by partitioning NMA into three neutral groups. As is apparent from a comparison with the standard CHARMM partial charges $Q_{\rm CHARMM}$ in the second data column of Tab. 1 our DFT/MM derived NMA model has a considerably larger dipole moment than a standard AG in CHARMM. We denote the thus obtained MM force field for NMA by MM(DFT). It becomes a PMM/II force field, if field dependent parameters $\Theta(\mathbf{E})$ are used for the harmonic bonding potentials (3) instead of the zero field parameters Θ^0.

4.3 Vibrations of the "AG within NMA"

With the NMA force field MM(DFT) at hand one could compute a Hessian $\mathbf{H}^0_{\rm MM(DFT)}$ and the corresponding NMA spectrum for comparison with results obtained from an MT/BP Hessian $\mathbf{H}^0_{\rm DFT}$. We are, however, solely interested in the force field of the amide core of NMA and in the corresponding MM parameters Θ^0. For a corresponding comparison the motions within NMA's methyl groups must be decoupled from those within the AG, such that pure amide modes can be extracted from the Hessians of NMA.

We accomplished the required decoupling by reducing the masses $m(\mathrm{H_{meth}})$ of the methyl hydrogens so strongly that the vibrational frequencies of the methyl modes become shifted well above the frequency of any other mode. A reduction by a factor of 10^{-6} turned out to render methyl frequencies that were at least two times larger than the frequency of the N—H stretch, which is the AG mode with the highest frequency. In the lower frequency part of the spectrum this mass-induced decoupling of the methyl modes yields a set of 12 normal modes. They are linear combinations of the 12 internal coordinates q_i defining an AG (cf. Fig. 3). Therefore they are called modes of the "AG within NMA". If these modes and their frequencies are calculated from two Hessians of NMA, a comparison indicates the similarity of the underlying force fields in the AG part of the molecule.

4.4 Fitting of linear response parameters

Assume that a force field parameter $p \in \Theta$ shows a linear response

$$p(\mathbf{E}) = p(0) + \boldsymbol{\alpha}\mathbf{E} \tag{6}$$

to an external electric field $\mathbf{E} = (E_x, E_y, E_z)^T$, where $p(0)$ is the zero-field value and $\boldsymbol{\alpha} = (\alpha_x, \alpha_y, \alpha_z)^T$ are the response parameters. For each of the field directions E_j, $j = x, y, z$, the parameter p may be an anti-symmetric or symmetric function of E_j. Accounting for this symmetry the component α_j of $\boldsymbol{\alpha}$ is given by

$$\alpha_j = \begin{cases} \alpha_j^+, & \text{if} \quad p(-E_j) = -p(E_j), \\ \mathrm{sgn}(E_j)\alpha_j^+, & \text{if} \quad p(-E_j) = p(E_j), \end{cases} \tag{7}$$

where α_j^+ is the linear response parameter at positive fields E_j. By the following steps one can determine α_j^+ and the zero-field parameter $p(0)$ from a data set of values $p(E_j)$, which has been derived from DFT results on NMA exposed to various constant electric fields E_x, E_y, and E_z.

First $p(E_j)$ is fitted[49] for each direction $j = x, y, z$ to a linear response function

$$p_j(E_j) = \begin{cases} p_j^0 + \beta_j^- E_j\,, & \text{if} \quad E_j < 0 \\ p_j^0 + \beta_j^+ E_j\,, & \text{otherwise.} \end{cases} \tag{8}$$

By averaging over the three zero-field estimates $p_j(0)$ we obtain a unique estimate $p(0)$ for the zero-field parameter. Inserting this value into Eq. (8) we repeat the fits and obtain slightly modified response parameters $\beta_j^\pm$. Taking the symmetry of $p(E_j)$ into account yields the desired response constant as the average value

$$\alpha_j^+ = \frac{1}{2}\left[\beta_j^+ + \mathrm{sgn}(\beta_j^+\beta_j^-)\beta_j^-\right]. \tag{9}$$

4.5 IR spectra from MD simulations of NMA

With the aim of computing IR spectra from Fourier transforms of time correlation functions (FTTCF, see Refs. 31,45), we carried out several MD simulations of NMA in the vacuum and in aqueous solution. For the vacuum simulations we employed EGO

with the NMA force field described in Sec. 4.2. For the simulations in D_2O a periodic rhombic dodecahedron with an inner radius R_i of 29 Å was filled with 4497 deuterated water models described by the Jorgensen's transferable four-point potential[50] (TIP4P) and a single NMA model. The system was equilibrated for several nanoseconds in the NpT ensemble using a Berendsen thermostat[51] (coupling time 0.5 ps, target temperature 293 K) and barostat[51] (coupling time 5 ps, isothermal compressibility $4.9\times10^{-10}\,\mathrm{Pa}^{-1}$, target pressure $1\times10^5\,\mathrm{Pa}$). The electrostatics were treated by toroidal boundary conditions,[38] which include a moving-boundary reaction field description for electrostatic interactions at distances larger than R_i and fast hierarchical multipole expansions up to the quadrupole moment.[52,53] The equations of motion were integrated with the Verlet algorithm[54] at a time step of 0.4 fs.

To obtain most accurate references for the quality of our PMM/II force field we have carried out two 85 ps MD simulations of the NMA/D_2O system in the NVT ensemble and in the DFT/MM setting.[9] In both simulations, NMA was chosen as the DFT fragment and was treated by the MT/BP approach (cf. Sec. 4.1). In one of these simulations NMA was protonated and in the other deuterated. Temperature was controlled by the above thermostat which was exclusively coupled to the solvent. To keep the NMA molecule near the center of the simulation system a harmonic potential with a force constant of 0.5 kcal/(mol Å^2) was applied to the molecule's center of mass. Considering the first 7 ps of the simulations as an adaptation to the DFT/MM force field we employed only the trajectories covering the remaining 78.6 ps for data analysis and for a FTTCF computation[31] of NMA's IR spectrum. In particular, we obtained the partial charges Q_{pol} of NMA listed in Tab. 1 from the trajectory of protonated NMA.

As a first test we additionally carried out PMM/II-MD simulations on protonated and deuterated NMA in D_2O. Here, we employed a substantially smaller system, which was also shaped as a periodic rhombic dodecahedron. But this time the inner radius was only $R_i = 19$ Å and it was filled with only 1403 deuterated TIP4P water models. The NMA model was initially described by our static MM(DFT) force field and the NpT equilibration procedure was adopted from the larger system. Subsequently, a short (50 ps) NVT equilibration preceded PMM/II production runs of twice the 78.6 ps duration, which was used in the DFT/MM setting. For a smooth sampling of the field-dependent force field parameters we employed a shorter integration time step of 0.2 fs.

In the PMM/II simulations the parameters $p_i[\mathbf{E}(t)]$ of the harmonic AG force field were updated together with the atomic coordinates $\mathbf{r}_i(t)$ at each integration time step t and were used in the following time step $t+\Delta t$ for the evaluation of the forces within the AG. The linear response relation Eq. (6) was applied to the calculation of the $p_i[\mathbf{E}(t)]$. Here $\mathbf{E}(t)$ represents a certain local average of the fields $\mathbf{E}[\mathbf{r}_i(t)]$ acting at the positions $\mathbf{r}_i(t)$ of the atoms i making up the AG. For the stretch potentials of the bonds C_α—C′ and N—C_α we used the averages of the fields $\mathbf{E}[\mathbf{r}_i(t)]$ at the respective bonded atoms, and for all other parameters we took the average $\langle\mathbf{E}(t)\rangle_c$ of the fields $\mathbf{E}[\mathbf{r}_i(t)]$ at the four atoms OC′NH making up the core of the AG.

As is well known, every MM-MD code calculates the fields $\mathbf{E}[\mathbf{r}_i(t)]$, which act on the partial charge Q_i of atom i and are generated by essentially all other partial charges in a simulation system, at each integration time step t. Here "essentially" means that electrostatic interactions of atoms joined through a few covalent bonds are (partially) neglected in many force fields. For the computation of the parameters $p_i[\mathbf{E}(t)]$ we extended the associated CHARMM22 convention[35] by the requirement that the partial charges within an AG cannot contribute to the polarization of the local force field. For the computation of the electrostatic forces $\mathbf{F}_i(t) = Q_i\,\mathbf{E}[\mathbf{r}_i(t)]$ acting on the partially charged atoms, however, we retained the CHARMM22 convention.

In the DFT/MM case, the IR spectra calculated by FTTCF for NMA were obtained from MD trajectories of its dipole moment $\mathbf{d}(t)$ covering $3\cdot2^{16}$ time steps with a duration of 0.4 fs, i.e. extending over a time span of 78.6 ps. The PMM/II simulations covered twice this time span. The autocorrelation function of $\mathbf{d}(t)$ was Fourier transformed and the resulting power spectrum was scaled by the so-called harmonic approximation quantum correction factor.[31,55] By Nyquist's theorem[56] the resulting spectral resolution is at least $0.4\,\mathrm{cm}^{-1}$ (DFT/MM). Noise reduction by convolution with a Gaussian kernel of $3\,\mathrm{cm}^{-1}$ standard deviation reduces the actual resolution of the FTTCF spectra accordingly. The bands in the FTTCF spectra were assigned by additional normal mode analyses. In the case of solution spectra we carried out instantaneous normal mode analyses following the procedures described

in Ref. 31. Gas phase line spectra were calculated by common normal mode analyses.

5 Results

As explained in Sec. 3.1, we will derive our PMM/II force field for AGs from DFT calculations on NMA. The quality of the results will depend on the accuracy at which the chosen DFT approach can model the intra-molecular force field of NMA. As a sensitive accuracy measure we will, therefore, first check the DFT description of NMA's gas-phase vibrational spectra.

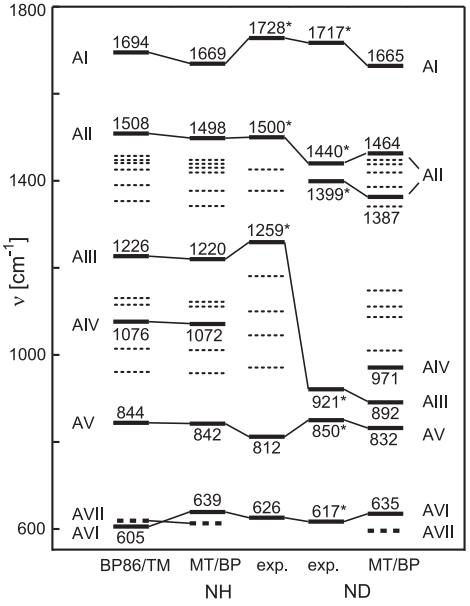

Figure 4: Comparison of experimental gas phase data with unscaled harmonic frequencies calculated by the DFT approaches BP86/TM and MT/BP for protonated (NH) and deuterated (ND) NMA in the spectral range 550 $\text{cm}^{-1} < \nu < 1800\ \text{cm}^{-1}$. The experimental data derive from IR[21] and resonance Raman[22] (marked by asterisks) measurements. Dotted lines indicate modes dominated by methyl motions, dashed lines out-of-plane amide modes, and solid lines in-plane amide modes, which are labeled AI to AVII in the order of decreasing MT/BP frequencies.

5.1 Evaluation of the MT/BP description

Figure 4 compares vibrational frequencies observed for protonated and deuterated NMA in the gas-phase[21,22] with unscaled harmonic frequencies calculated by the DFT approaches BP86/TM and MT/BP introduced in Sec. 4.1. Modes involving in-plane (ip) motions of AG atoms are emphasized by solid lines and are labeled AI to AVI in descending order of the MT/BP-NH results. The amide mode A0 with the highest frequency, the N—H stretching mode (MT/BP: 3556 cm^{-1}), is outside the depicted frequency range $[550, 1800]\ \text{cm}^{-1}$, the highest frequency out-of-plane (oop) mode is near the lower border of that range (dashed line, AVII).

A look at the first two columns in Fig. 4 immediately shows that the BP86/TM and MT/BP descriptions of the NMA force field agree quite well, as was to be expected from previous comparisons of this kind for other molecules.[45,57] For the frequencies of the amide modes AI-AIV, e.g., the root mean square deviation is 18 cm^{-1}. In the high-frequency range ($\nu > 1000\ \text{cm}^{-1}$) the NMA force field is slightly softer for MT/BP than for BP86/TM. The mutual match of calculated normal modes is even more convincing (data not shown). Thus, the differences between the two BP86 descriptions, one employing a large Gaussian basis set for all the electrons (BP86/TM), and the other one employing a large plane wave basis set for the valence electrons combined with pseudo-potentials (MT/BP) are quite small in the given case of an isolated NMA molecule.

To check whether this agreement prevails when NMA is exposed to external electric fields we have calculated the NMA vibrational spectra for a homogeneous electric field $E_y = 10\ \text{kcal}/(\text{mol}\,\text{Å}\,\text{e})$ oriented along the y-axis of the coordinate system depicted in Fig. 2. We found a very close agreement for all field-induced frequency shifts obtained from the two computational methods (data not shown). For instance, the red-shifts of the AI frequency (expected for such a field from the discussion of the resonance structures in Fig. 2) were 27 cm^{-1} for BP86/TM and 29 cm^{-1} for MT/BP.

As was also to be expected from previous BP86 studies of vibrational spectra,[6,7,45] the calculated harmonic frequencies agree quite well with the observed fundamentals. Comparing in Fig. 4 for protonated NMA the assigned (thin lines) theoretical and experimental frequencies (columns 1-3) one finds root mean square deviations of only 35 cm^{-1}

for MT/BP and 27 cm^{-1} for BP86/TM. Here, BP86/TM matches the experimental data slightly better than MT/BP, because in the high-frequency region the associated force fields become softer in the sequence: "exp." $\rightarrow$ BP86/TM $\rightarrow$ MT/BP. Correspondingly, in this important spectral region both DFT descriptions slightly underestimate the amide frequencies.

Because MT/BP underestimates the frequencies of the high-frequency amide modes AI-AIII by up to 3.4% the method correspondingly underestimates the isotope shifts observed upon deuteration of the amide nitrogen (compare columns 2-5 in Fig. 4). For instance, the very large (338 cm^{-1}) isotope shift of the AIII mode, which mainly belongs to the ip bending motion of the amide hydrogen, is underestimated by 3%. This observation indicates that a scaling of MT/BP frequencies, which leads to a closer match with experimental data for unsubstituted NMA, will also improve the agreement for isotopically substituted variants.

Table 2: Mode specific scaling factors for NMA.

mode	$\nu^{\mathrm{NH}}_{\mathrm{exp}}/\nu^{\mathrm{NH}}_{\mathrm{MT/BP}}$	$\nu^{\mathrm{ND}}_{\mathrm{exp}}/\nu^{\mathrm{ND}}_{\mathrm{MT/BP}}$
AI	1.0354	1.0312
AII	1.0013	0.9958
AIII	1.0320	1.0325
AV	0.9644	1.0216
AVI	0.9797	0.9717

The data collected in Tab. 2 furthermore suggest that the scaling of the MT/BP frequencies should be preferentially chosen mode specific. The table lists AI-AVI frequency ratios $\nu_{\mathrm{exp}}/\nu_{\mathrm{MT/BP}}$ for NMA's amide modes, which have been calculated from the data in Fig. 4 (for the two AII modes present in the ND spectra we have chosen the average frequency). With the exception of the mode AV the frequency ratios are seen to be nearly invariant upon deuteration indicating that a mode specific scaling should actually make sense.

5.2 Harmonic zero-field parameters

We conclude from the results presented in the preceding paragraph that MT/BP calculations on NMA should offer a reasonable starting point for the development of a PMM/II force field for AGs. Because the MT/BP spectrum of NMA shown in Fig. 4 was derived from a corresponding equilibrium geometry $\mathcal{G}^0$ and Hessian $\mathbf{H}^0$ at zero field, these data ($\mathcal{G}^0$, $\mathbf{H}^0$) immediately lead to a first set Θ^0 of harmonic zero-field parameters. As explained in Secs. 3.2 and 4.2 the harmonic model potentials Eq. (3) and Eq. (4) belonging to the 12 internal coordinates q_j depicted in Fig. 3 are specified by Θ^0. The equilibrium values q_j^0 of the q_j are directly given by $\mathcal{G}^0$ and the harmonic force constants k_j follow from $\mathbf{H}^0$ by the iterative procedure outlined in Sec. 3.2.

The resulting parameter set is listed in Tab. 3 and is called "MM(DFT)", because the force constants k_i (except for $i = 12$) have been directly derived from the MT/BP Hessian $\mathbf{H}^0$. Note that according to the values displayed for the k_i of the two improper dihedral angles ϕ_1 and ϕ_2 (cf. Fig. 3) the sp^2 hybridization is much more rigidly defined at the C′ atom than at the amide nitrogen. For the harmonic force constant $k_{12} = k_\xi$, which can be associated to the proper dihedral angle ξ, the MT/BP Hessian yields a value of 0.3 kcal/{mol(10deg)^2}. According to the arguments in Sec. 3.2 this value does not suffice to determine the energies E_1 and E_2, which parameterize the dihedral potential (2). For a complete specification one additionally needs the *cis-trans* energy difference E_1, which is easily accessible by DFT:

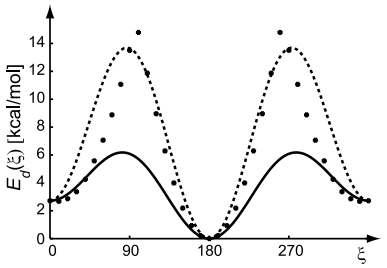

Figure 5: *Cis-trans* torsional potential curves $E_d(\xi)$ calculated by BP86/TM (dots) and by the model potential Eq. (2) using for the *cis-trans* energy difference E_1 the BP86/TM value of 2.7 kcal/mol and for E_2 the values 4.7 kcal/mol (solid line) and 12.3 kcal/mol (dashed line), respectively. See the text for a discussion.

Figure 5 shows the results of a BP86/TM calculation on the *cis-trans* torsional potential curve $E_d(\xi)$, according to which E_1 is 2.72 kcal/mol. Combined with the above value of k_ξ one gets $E_2 =$ 4.7 kcal/mol. In Fig. 5 the corresponding model potential curve Eq. (2) is drawn as a solid line. This curve closely approximates the BP86/TM data near the *trans* state at $\xi = 180$ deg but strongly

Table 3: Harmonic zero-field parameters MM(DFT) for an AG.

i	q_i	q_i^0	k_i	i	q_i	q_i^0	k_i
1	l_1	1.520 Å	460*	7	ψ_2	122.6 deg	6.9$
2	l_2	1.233 Å	1443*	8	ψ_3	118.5 deg	4.2$
3	l_3	1.374 Å	748*	9	ψ_4	122.5 deg	4.9$
4	l_4	1.015 Å	1009*	10	ϕ_1	0 deg	7.1$
5	l_5	1.455 Å	591*	11	ϕ_2	0 deg	1.2$
6	ψ_1	121.9 deg	4.7$	12	ξ	180 deg	0.8$

*[kcal/(mol Å^2)], $[kcal/{mol (10 deg)^2}].

underestimates the torsion barrier. A better estimate for this barrier is obtained (dashed line), if E_2 is increased to 12.3 kcal/mol. The corresponding model potential is substantially stiffer near $\xi = 180$ deg as is demonstrated by the value $k_\xi = 0.8\,\text{kcal}/\{\text{mol}\,(10\,\text{deg})^2\}$ of the associated harmonic force constant. Because of the better estimate for the torsional barrier this latter value has been included into the parameter set MM(DFT) listed in Tab. 3.

Note that the dihedral angles ϕ_1, ϕ_2, and ξ play a special role in the MM force field of AGs: In a planar AG the ip and oop vibrations are strictly decoupled and the three dihedral angles exclusively contribute to three oop modes found at frequencies well below 800 cm^{-1} (cf. Fig. 4). Because we are mainly interested in the ip modes at frequencies above 1000 cm^{-1}, our AG force field can tolerate large inaccuracies in the model potentials of the dihedrals as long as it produces during an MD simulation largely planar AG structures. Therefore, the choice of a stiffer torsion potential (dashed line in Fig. 5) is advantageous.

5.3 Quality of the harmonic model at zero field

Once the parameter set MM(DFT) is given, the associated AG force field Eq. (1) can be easily extended towards an MM force field for the whole NMA molecule by adding standard parameters for the attached methyl groups (cf. Sec. 4.2). By computing NMA's equilibrium geometry and Hessian with this force field one can compute an MM estimate for the gas-phase spectrum of NMA. However, such a spectrum will not only carry the signatures of the AG parameters derived from DFT but also those of the standard MM model chosen for NMA's methyl groups. Due to the mixing of AG modes with methyl modes a comparison of MT/BP or experimental NMA spectra with such an MM spectrum will not easily render clues on the quality of the AG force field derived from DFT. To concentrate such a quality estimate on the "AG within NMA" we have designed a procedure for the calculation of corresponding spectra that is based on the use of essentially massless hydrogen atoms within NMA's methyl groups (cf. Sec. 4.3).

Figure 6 compares the spectrum of the "AG within NMA" calculated from the force field MM(DFT) with the MT/BP reference and with a hypothetical "experimental" spectrum, which was estimated from the MT/BP frequencies as explained in the caption. Almost all MM(DFT) frequencies are seen to differ significantly from the MT/BP and "experimental" references. However, a comparison of the associated normal modes depicted in Fig. 7 demonstrates that the mode compositions are extremely similar.

Optimization attempts, which considered the harmonic force constants as free parameters and aimed at an improved matching of frequencies, have shown that within the given MM model [Eqs. (1)-(3)] improved frequencies are inevitably connected with a strong distortion of the normal modes (data presented in the supporting material: Figs. 17, 18, Tab. 5). These results indicate that a better MM description of frequencies, which preserves the good performance on the normal modes, can only be obtained by including at least a few of the coupling terms (5) into the model (see also Fig. 14 in the supporting information, which compares DFT and MM(DFT) Hessians). Thus, in future improvement efforts one will have to determine which of these coupling terms are most important (see Ref. 37 for possible methods). But whatever the final choice may be, it needs to be extended by a description of the field-dependence, if it is supposed to serve as the core of a PMM/II force field for AGs.

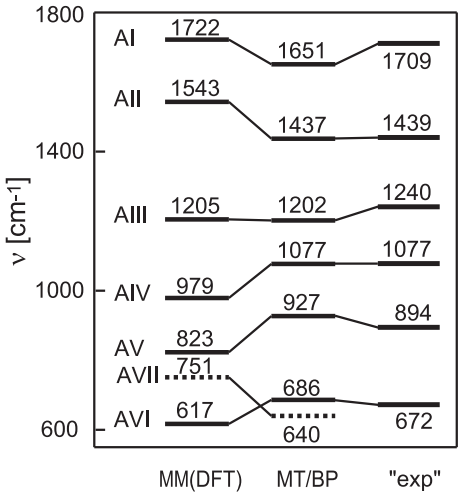

Figure 6: Spectra of the "AG within NMA" (defined by methyl groups with nearly massless hydrogen atoms, cf. Sec. 4.3) obtained from the harmonic force field MM(DFT) and from MT/BP. A hypothetical "experimental" spectrum ("exp") was estimated by scaling the MT/BP results with the factors listed in the first data column of Tab. 2 (for amide modes, to which we could not assign a band in the experimental NMA spectrum, we chose the factor one). For line styles and mode labels see the caption to Fig. 4.

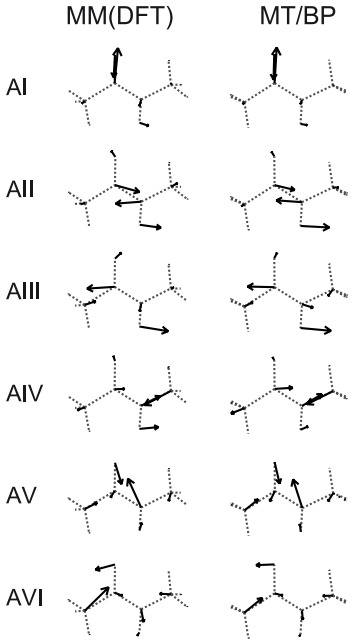

Figure 7: In-plane normal modes of the "AG within NMA" (cf. Sec. 4.3) belonging to the line spectra depicted in Fig. 6. Dashed lines indicate the bonds within the NMA molecule covering the AG depicted in Fig. 3. Arrows visualize atomic motions.

5.4 Field dependence

As outlined in Secs. 3.1 and 3.2 we have calculated the MT/BP equilibrium geometries $\mathcal{G}(\mathbf{E})$ and Hessians $\mathbf{H}(\mathbf{E})$ of NMA exposed to homogeneous external fields of varying strengths and directions. These fields were chosen parallel to the Cartesian coordinate axes depicted in Fig. 2 with values $E_j \in [-30, +30]$ kcal/(mol Å e), $j \in \{x, y, z\}$. The chosen range covers typical field strengths occurring in condensed phase. For instance in our MD simulations of protonated and deuterated NMA in heavy TIP4P water (cf. Sec. 4.5) the average external field $\langle \mathbf{E} \rangle$ acting on NMA turned out to be oriented mainly along the negative y axis and, at the four atoms O=C′–N–H making up the core of NMA's AG, had the average strength $\langle |\mathbf{E}| \rangle = 34$ kcal/(mol Å e). In SI units this is about 15×10^9 V/m.

The MT/BP results $[\mathcal{G}(E_j),\ \mathbf{H}(E_j)]$ were converted into field-dependent parameter sets $\Theta(E_j)$ of the harmonic AG force field (3) by the procedures explained in Sec. 3.2. The sets $\Theta(E_j)$ comprise the equilibrium values $q_i^0(E_j)$ and force constants $k_i(E_j)$ associated to the nine internal coordinates q_i describing the bond lengths l_i and bond angles ψ_i of an AG. They furthermore contain the force constants $k_i(E_j)$ belonging to the two improper dihedral angles ϕ_i (the associated ϕ_i^0 were set to zero forcing the AG to be planar). As mentioned above we chose the torsion potential $E_d(\xi)$ associated to the proper dihedral angle ξ as constant, because BP86/TM results indicated the field-dependence to be small (data not shown).

The data sets $p_i(E_j)$ on the field-dependence of the force field parameters $p_i \in \{q_i^0, k_i\}$ were visualized by a series of graphs depicting for each field direction j the relative field-induced changes $p_i(E_j)/p_i(0)$ as a function of E_j (see Figs. 12 and 13 in the supplementary material). As a typical example Figure 8 shows such graphs for the equilibrium lengths $l_i^0(E_y)$ and force constants $k_i^0(E_y)$ of the C′=O and C′—N bonds. We have chosen this example, because the important AI and AII modes of the AGs are dominated by the correspond-

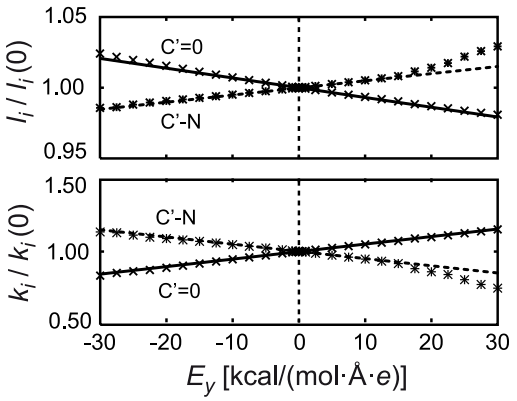

Figure 8: Equilibrium lengths $l_i/l_i(0)$ and force constants $k_i/k_i(0)$ of the C′=O ($i = 2$) and C′—N ($i = 3$) bonds of NMA for different electric fields E_y oriented along the y-axis of the coordinate system in Fig. 2. Crosses mark MT/BP results, lines are linear fits for $E_y \in [-10, 10]$ kcal/(mol Å e).

ing stretching vibrations and because for NMA in a polar solvent the average reaction field is oriented along the negative y-direction.

According to Fig. 8 the C′=O bond becomes longer and its force constant softer for values $E_y < 0$. At a field $E_y = -30$ kcal/(mol Å e), which is typical for an aqueous environment, the elongation is about 2.5 % and the softening about 20 %. This behavior confirms and quantifies the corresponding expectations derived by us in Sec. 2.1 from the π-resonance structures depicted in Fig. 2. In line with our expectations the C′—N bond shows the opposite behavior with slightly smaller relative changes.

The most important aspects of the data depicted in Fig. 8 (and in the supplementary Figs. 12 and 13) and the key results of this paper are (i) the considerable sizes of the changes, which are induced by typical external fields into the force constants of the O=C′–N–H core of an AG and (ii) their nearly linear field dependences covering wide ranges of field strengths.

Concerning (i) the following estimate illustrates that these purely electrostatic effects of polarization are strong enough to completely explain e.g. the solvatochromic shifts observed for NMA in polar solvents: If the frequency of the AI band was exclusively determined by the force constant $k_{C'O}$, one would expect from Fig. 8 and $\nu \propto \sqrt{k}$ the AI frequency to shift from 1730 to 1596 cm^{-1} upon transfer of NMA from the gas-phase into an aqueous environment. Experimentally, a smaller shift to 1625 cm^{-1} is observed.[20] This diminished red-shift is caused by the coupling of the AI with the lower-frequency AII mode. Assigning the latter exclusively to $k_{C'N}$ and applying the same simplistic reasoning once again, one predicts a blue-shift of the AII band of NMA from 1500 to 1595 cm^{-1}. Again, the coupling to the higher-frequency AI mode, which is missing in our simple estimate, explains the slightly lower experimental value[20] of 1582 cm^{-1}.

Consequently, the response data depicted in Fig. 8 nourish the hope that a PMM/II approach can lead to quantitative descriptions of amide bands. Concurrently the largely linear response will enable a simple implementation. Only at large fields deviations from linear response are observed, part of which (see e.g. the behavior of $l_{C'N}(E_y)$ for $E_y \gg 0$) may represent computational artifacts.

As documented by the set of graphs in the supplementary material (see Figs. 12 and 13), the MT/BP calculations predict at small fields, i.e. for $|\mathbf{E}| \leq 10$ kcal/(mol Å e), a linear response behavior for all force field parameters p_i with the notable exception of the force constant k_{11} of the improper dihedral angle ϕ_2 measuring the stability of the sp^2 hybridization at the amide nitrogen. For k_{11} the field dependence is much larger than for any other parameter with the data indicating that the sp^2 hybridization is strongly stabilized in condensed phase. However, these results are hampered both by non-linearities and computational artifacts. On the other hand, the associated model potential solely contributes to low-frequency oop vibrations and, thus, inaccuracies in the corresponding parameter estimates can be presently tolerated.

Applying the fit-procedure described in Sec. 4.4 to the MT/BP derived data sets $p_i(E_j)$ at small electric fields $E_j \in [-10, 10]$ kcal/(mol Å e), yields for each parameter p_i a linear response constant $\boldsymbol{\alpha}_i$ [including cases in which p_i turned out to be a symmetric function $p_i(E_j) = p_i(-E_j)$]. Table 4 lists the resulting linear response parameters $\boldsymbol{\alpha}_i = (\alpha_{i,x}, \alpha_{i,y}, \alpha_{i,z})^T$ normalized to the zero-field value $p_i(0)$ of the associated force field parameter and marks symmetric functions by an asterisk. The data in Tab. 4 combined with the harmonic zero-field parameters MM(DFT) listed in Tab. 3 represent the core of our suggestion for a first most simple PMM/II force field for AGs.

Table 4: Field dependence[$] of force field parameters $p_i \in \{q_i^0, k_i\}$.

i	$\alpha^+_{i,x}/q_i^0(0)$	$\alpha^+_{i,y}/q_i^0(0)$	$\alpha^+_{i,z}/q_i^0(0)$	$\alpha^+_{i,x}/k_i(0)$	$\alpha^+_{i,y}/k_i(0)$	$\alpha^+_{i,z}/k_i(0)$
1	0.51	0.21	-0.05^*	-4.87	-2.60	0.32
2	0.20	-0.68	0.08	-1.69	5.24	0.83^*
3	-0.87	0.50	0.39^*	9.14	-5.04	-3.65^*
4	0.03	-0.10	0.14^*	-0.14^*	-0.35^*	-1.03^*
5	0.44	0.12	0.21^*	-3.39	-1.16	-1.47^*
6	-0.92	-0.10	0.11^*	-3.99	-3.82	0.33^*
7	0.94	-0.02	-0.02^*	-3.74	-6.84	-2.07^*
8	0.10	-0.09^*	-1.16^*	1.01	-2.62	0.59^*
9	0.63	-0.52	-0.75^*	-3.22	1.11	2.72^*
10	—	—	—	-1.92	0.28^*	0.61^*
11	—	—	—	12.84	-29.42	32.19^*

[$]measured by the parameters $\boldsymbol{\alpha}_i^+/p_i(0)$ of relative linear response which are given in units of $(\mathrm{kmol\,\AA\,}e)/\mathrm{kcal}$

[*]parameters p_i with $p_i(-E_j) = p_i(E_j)$, $j = x, y, z$.

5.5 Proof of principle

Quality checks for a new force field represent yet another challenge. One has to have access to reliable experimental and high-quality computational reference data. For force fields aiming at the description of vibrational spectra the reference data sets should preferentially cover many isotopomers of relevant model compounds. Unfortunately such an extended data base does not yet exist for AGs. Therefore we restrict the first checks concerning the performance of our newly designed PMM/II force field to the case of the NMA model compound.

To provide a high quality computational reference for NMA's condensed phase IR spectra we have carried out extended DFT/MM simulations for the protonated and deuterated isotopomers of this molecule in aqueous solution. Here, NMA was described by the MT/BP approach employed for force field derivation and the aqueous environment by about 4500 deuterated TIP4P water molecules. The dimensions of the periodic simulation system were chosen such that it models an aqueous solution at room temperature and ambient pressure as closely as possible (see Sec. 4.5 for details). The DFT/MM simulations yielded two 85 ps trajectories of NMA's dipole moment $\mathbf{d}(t)$ sampled at time steps of 0.4 fs from which the last 78.6 ps were selected for data analysis. Each of the two simulations covered as the most time consuming steps 212500 MT/BP calculations on NMA and required about 160 days computing time on a cluster of eight 2.4 GHz processors.

The MT/BP solution spectrum of protonated NMA depicted in Fig. 9C was obtained from $\mathbf{d}(t)$ through the FTTCF approach described in Sec. 4.5 and by scaling the frequencies homogeneously with a factor of 1.0354. This factor has also been applied to the MT/BP gas-phase line spectrum shown in Fig. 9B. It has been chosen such that, for the gas phase, the MT/BP prediction of the AI frequency matches the resonance Raman value[22] of 1728 cm^{-1} (cf. Fig. 4).

As is apparent from a comparison with the published[20] IR spectra reproduced in Figs. 9A (gas phase) and 9D (aqueous solution) the scaled DFT and DFT/MM results in Figs. 9B and 9C are actually "high quality" references for judging corresponding PMM/II descriptions. Particularly the DFT/MM result for NMA in aqueous solution (Fig. 9C) can nicely explain and reproduce all corresponding observed (Fig. 9D) spectral features including those in the congested spectral region between 1300 and 1500 cm^{-1}, which predominantly contains normal modes localized in NMA's methyl groups. The quality, at which such a FTTCF spectrum derived from DFT/MM-MD can account for the experimental data, not only depends on the chosen DFT/MM hybrid model but also on the duration of the MD simulation, i.e. on how well the equilibrium fluctuations were sampled. Fig. 15A in the supporting material provides corresponding information. Note furthermore that the mode assignments in Fig. 9C have been obtained from a series of DFT/MM instantaneous normal mode analyses ap-

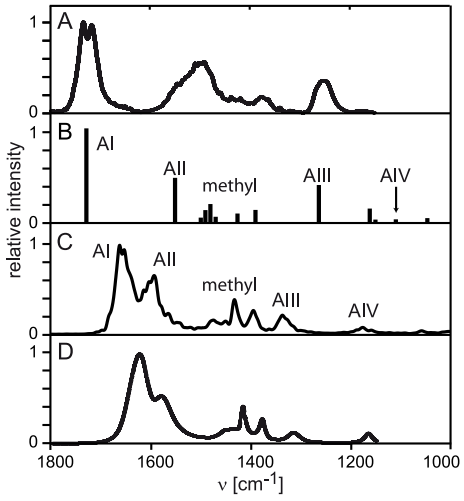

Figure 9: Experimental IR spectra[20] (originally published by Kubelka and Keiderling in J. Phys. Chem., 2001) of protonated NMA in the gas phase (A) and in aqueous solution (D) are compared with MT/BP and MT/BP/MM-MD descriptions obtained by normal mode analyis for the gas phase (B) and by FTTCF for D_2O (C), respectively. All intensities are normalized to the respective maximal values. As explained in the text the DFT and DFT/MM frequencies were scaled by a factor of 1.0354. Data files of the experimental spectra were kindly provided by Jan Kubelka.

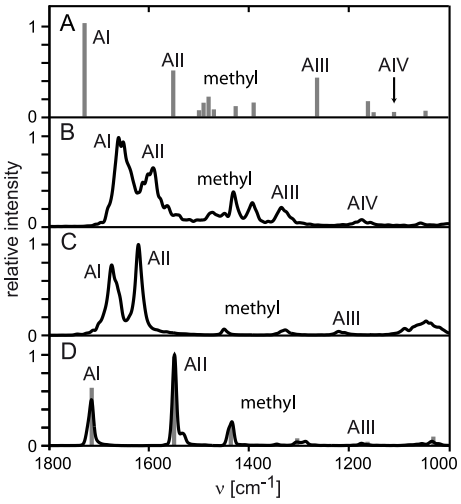

Figure 10: IR spectra calculated for protonated NMA in the gas phase (A,D) and in aqueous solution (B,C). Gray bars denote line spectra from normal mode analyses of isolated NMA: (A) MT/BP and (D) PMM/II descriptions. Solid lines are FTTCF results from extended MD trajectories: DFT/MM-MD (B) and PMM/II-MD (C) of NMA in heavy TIP4P water; (D) PMM/II-MD of isolated NMA. Intensities are normalized to the respective maximal values and frequencies have been scaled. For explanation and details see the text.

plied to snapshots of the DFT/MM-MD trajectory (cf. Sec. 4.5).

Having assured the quality of the reference spectra in Figs. 9B and particularly 9C we can now turn to the evaluation of the corresponding PMM/II description. For the purpose of a most simple visual comparison we have reproduced in Figs. 10A and 10B once again the DFT and DFT/MM reference spectra. In Figs. 10C and 10D we show the corresponding PMM/II results.

Clearly, the PMM/II gas phase spectrum in Fig. 10D cannot give any insight into the quality, at which the new force field models NMA's response properties, because in this case there is no external field. Instead this spectrum shows how the zero-field model MM(DFT) for NMA, which combines a standard CHARMM description of the methyl groups and the harmonic force field Eq. (1) for the AG with a simplified charge distribution within NMA (Tab. 1), performs on the IR spectrum of the isolated NMA molecule.

Further above, in connection with the discussion of Fig. 6, we have already carried out a similar evaluation of the force field MM(DFT) in which we, however, had largely excluded the influence of the methyl parameterization by concentrating on the "AG within NMA". As is apparent from Fig. 6 for the "AG within NMA" the MM(DFT) force field overestimates the frequencies of the AI and AII modes while underestimating the AIII frequency. Scaling of the MM(DFT) frequencies by a factor of 0.9925 reduces this AI result from $1722\,\mathrm{cm}^{-1}$ to the hypothetical "experimental value" of $1709\,\mathrm{cm}^{-1}$.

The PMM/II spectra in Figs. 10C and 10D were scaled with this factor and the remaining difference between the experimental ($1728\,\mathrm{cm}^{-1}$) and MM(DFT) ($1716\,\mathrm{cm}^{-1}$) frequencies of NMA's AI mode in the gas phase indicate the influence of the methyl parameterization on the spectral location of this mode. When comparing the MM(DFT) inten-

sities in Fig. 10D with the MT/BP reference intensities in Fig. 10A one immediately recognizes large differences. For instance according to MM(DFT) the AII band should be much more intense than the AI band and the methyl bands should have very small intensities. These ill-descriptions are obvious consequences of our simplistic choice of the partial charges within NMA (cf. Tab. 1) and are of no concern in the present context.

A more important aspect of the MM(DFT) results in Fig. 10D on the isolated NMA molecule is the shown comparison of a line spectrum from normal mode analysis (gray bars) with an FTTCF spectrum from a 78.6 ps vacuum MD trajectory (solid line). Quite obviously the two methods give essentially identical frequencies and intensities thus verifying our FTTCF approach. The most important aspect of Fig. 10D, however, is that it enables a direct visual comparison with the PMM/II solution spectrum in Fig. 10C. For the same reason the DFT result for an isolated NMA molecule (Fig. 10A) has been drawn directly above the corresponding DFT/MM solution spectrum (Fig. 10B).

Comparing visually Fig. 10D with Fig. 10C (and Fig. 10A with Fig. 10B) reveals how the effects of solvation are covered by PMM/II (and by the DFT/MM reference, respectively). Quite apparently the description of the solvatochromic shifts, which is provided by the polarizable force field, is very similar to that delivered by the reference approach: According to both descriptions the AI band is shifted to the red and the AII and AIII bands are shifted to the blue upon transfer of NMA from the gas phase into an aqueous environment. Qualitatively, these effects are the same as those observed experimentally (compare Fig. 9A with Fig. 9D).

If we take the solvatochromic reduction $\Delta\Delta\nu_{solv} = \Delta\nu^{\rm gas}_{\rm AI/AII} - \Delta\nu^{\rm D_2O}_{\rm AI/AII}$ of the spectral gap $\Delta\nu_{\rm AI/AII}$ between the AI and AII bands as a quantitative measure, the computational reference method yields $\Delta\Delta\nu^{\rm DFT/MM}_{solv} = 112\,{\rm cm}^{-1}$ whereas PMM/II gives $\Delta\Delta\nu^{\rm PMM/II}_{solv} = 113\,{\rm cm}^{-1}$. These values have to be compared with the somewhat larger solvatochromic gap reduction $\Delta\Delta\nu^{\rm exp}_{solv} \approx 185\,{\rm cm}^{-1}$ observed experimentally. However, the latter gap reduction bears certain uncertainties due to the broad shape of the gas phase AII band (see Fig. 9A) and the possibility that the studied NMA vapor[20] contained sizable amounts of *cis*-NMA. According to our DFT descriptions the double peak displayed by Fig. 9A for the gas-phase AI band could be explained by a *cis*-NMA admixture. If we therefore take the computational reference data as our main measure for the quality at which the PMM/II approach can account for solvatochromic effects, we come to the conclusion that in this respect PMM/II performs quite well.

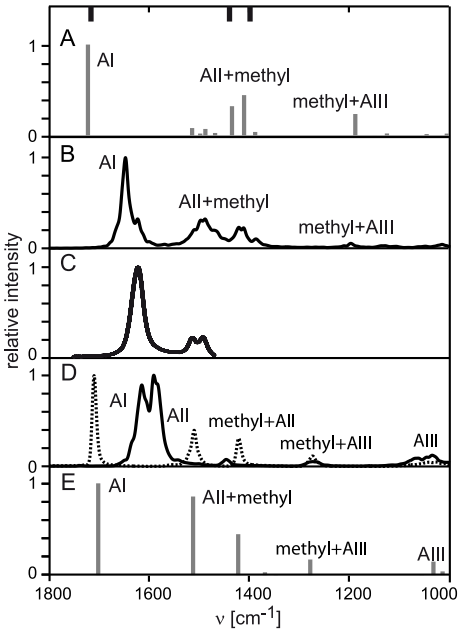

Figure 11: An experimental IR spectrum[20] (C) of deuterated NMA in D_2O is compared with corresponding spectra calculated for D_2O (B,D) and for the gas phase (A,E). For the gas phase (A,E) band positions observed by resonance raman spectroscopy[22] are indicated as thick black bars in the upper parts of the figures. Gray bars in (A,E) denote line spectra from normal mode analyses of the isolated molecule: (A) MT/BP, (E) PMM/II. Solid lines are FTTCF results from extended MD trajectories: DFT/MM-MD (B) and PMM/II-MD (D) of deuterated NMA in heavy TIP4P water (cf. the caption to Fig. 10). The additional dashed line in (D) is from a MM-MD simulation of NMA in TIP4P water with NMA described by the nonpolarizable force field MM(DFT).

As we have seen in the discussion of Fig. 4 the deuteration of the amide nitrogen decouples the AII and AIII modes of NMA by shifting the AIII frequency substantially to the red. As a result the AII mode of an isolated NMA molecule becomes an almost pure C'—N stretching vibration. On the other hand, the sizable field dependences of the C'=O and C'—N stretching force constants (cf. Fig. 8) have

suggested that the observed solvatochromic shifts should be mainly due to these field dependences. Therefore computations of the solvatochromic effects for deuterated NMA molecules can provide additional insights.

Fig. 11 compares the IR spectra of deuterated NMA obtained by several different methods (see the caption for explanation). Like in the protonated case we can also here take the solvatochromic reduction $\Delta\Delta\nu_{solv}$ of the AI/AII spectral gap as a measure. In this case the experimental value[20,22] $\Delta\Delta\nu_{solv}^{\mathrm{exp}}$ can be given more accurately, because the FTTCF spectrum depicted in Fig. 11B, which was obtained from the DFT/MM-MD simulation and has been assigned through a series of DFT/MM instantaneous normal mode analyses, uniquely enables the assignment of the small double peak centered at $1503\,\mathrm{cm}^{-1}$ in the experimental IR spectrum[20] (Fig. 11C) to the AII mode. With this new assignment and the clear (gas phase) resonance Raman data in Ref. 22 we get $\Delta\Delta\nu_{solv}^{\mathrm{exp}} \approx 157\,\mathrm{cm}^{-1}$, which is to be compared with the values $\Delta\Delta\nu_{solv}^{\mathrm{DFT/MM}} = 137\,\mathrm{cm}^{-1}$ and $\Delta\Delta\nu_{solv}^{\mathrm{PMM/II}} = 166\,\mathrm{cm}^{-1}$. In view of the difficulty to pin down exact frequencies for the AII mode, which frequently contributes to several combinations with methyl vibrations, all these values roughly agree with each other. Thus the good performance of PMM/II concerning the description of solvatochromic shifts stated above has become corroborated once again.

However, a closer look at the relative spectral positions of the AI and AII bands in Fig. 11D (solid line) reveals that the PMM/II force field must be further improved. Already in the zero field case one finds $\Delta\nu_{\mathrm{AI/AII}}^{\mathrm{gas,PMM/II}} = 190\,\mathrm{cm}^{-1}$, whereas scaled DFT predicts $288\,\mathrm{cm}^{-1}$ and experimentally one finds $277\,\mathrm{cm}^{-1}$ (cf. Figs. 11A and 11E). Thus, at zero field PMM/II strongly underestimates the AI/AII spectral gap. Due to the large ($166\,\mathrm{cm}^{-1}$) solvatochromic reduction of this gap the AI and AII bands strongly overlap in the PMM/II simulation of the aqueous solution. Correspondingly, instantaneous normal mode analyses revealed for PMM/II a mixing of the C'=O and C'—N stretches, whereas no such mixing was found for DFT/MM (FTTCF spectrum in Fig. 11B, convergence in Fig. 15B). This finding supports the conclusions of Sec. 5.2, according to which the purely harmonic AG model potential (3) needs to be extended by coupling terms (5) for a quantitative description of frequencies. Furthermore, it underlines our critique of the extended TDC methods (cf. Sec. 2.2) which argued with the necessity of correctly reproducing the frequency spacing among the AI-AIII modes.

To illustrate the difference between a conventional MM force field and a PMM/II approach we have included in Fig. 11D an additional FTTCF spectrum (dashed line) that was obtained from a MD simulation of deuterated NMA (described by the non-polarizable force field MM(DFT)) in a D_2O solution (modeled by TIP4P). This spectrum is nearly, but not completely, identical to the gas-phase spectrum (11E). The main differences are the reduced intensity of the AII band and a slight blue shift of the AI frequency. This blue shift is a consequence of hydrogen bonding between NMA's carbonyl group and the surrounding water molecules. In the protonated case (see Fig. 16 in the supporting material) also the AII mode undergoes a blue shift, because here it is coupled with the ip bending mode of the hydrogen bonded N—H group. In this case the solvatochromic blue shifts of the AII and AIII modes predicted according to Fig. 10 by DFT/MM-MD and PMM/II-MD for protonated NMA have a small hydrogen bonding contribution. For deuterated NMA, however, all solvatochromic effects are exclusively caused by polarization.

The stated good performance of PMM/II concerning solvatochromic shifts in the NMA spectra represents a proof of principle that this new type of polarizable force field for AGs should be capable of covering the key electrostatic effects which steer the shapes of protein and peptide amide bands. This proof of principle makes clear that a reliable PMM/II approach is now within reach. While similarly accounting for the effects of external electric fields the PMM/II approach is computationally by several orders of magnitude more efficient than the DFT/MM reference. In the case of NMA in solution the computational speed-up turned out to be a factor of 400. Instead of having to wait 160 days for a computed spectrum, such a spectrum becomes available within a few hours. Note in this connection that the computational effort required for computing protein or peptide spectra by PMM/II is of the same order of magnitude as the one spent by us on NMA in solution. The reasons are that a MD simulation with a PMM/II force field requires essentially the same computational effort as a classical MM-MD simulation, that the NMA/D_2O simulation system treated by us is not much smaller than simulation systems usually employed for proteins or peptides in solution, and that for the PMM/II computation of an IR spectrum the simulation time

does not have to exceed a few ten or hundred picoseconds.

6 Summary and outlook

The above suggestion of a PMM/II force field for AGs represents an important milestone on the way towards computationally efficient and nevertheless sufficiently accurate descriptions of the IR spectra of polypeptides. Up to now our suggestion solely covers a set of largely harmonic energy functions associated to a non-redundant set of internal coordinates for AGs, various sets of zero-field parameters for these functions, and for each parameter involved in one of the harmonic potentials a linear response function. The core set of these parameters, which comprises the values listed in Tab. 3 defining the AG force field "MM(DFT)" and the linear response constants in Tab. 4, was derived from DFT calculations on the model compound NMA using the MT/BP approach. These DFT calculations demonstrated that linear response actually applies to most of these parameters even at the strong fields prevalent in condensed phase environments. They furthermore showed that such fields can change the force constants of the internal coordinates in an amide group by 20 % or more. Changes of this size suffice to explain, e.g., the variations of the amide I frequencies in different protein structure or solvent environments.

For an isolated AG the MM(DFT) force field turned out to nicely reproduce the MT/BP normal mode compositions while exhibiting certain deviations of calculated frequencies. For protonated and deuterated NMA in aqueous solution, for which DFT/MM simulations were shown to yield excellent descriptions of the observed IR spectra, the large solvatochromic effects were seen to be quite well covered by the much simpler PMM/II approach (despite the retained frequency deviations). This result was taken by us as a "proof of principle" that the PMM/II approach works.

However, with the achievements sketched above the development of a PMM/II force field for AGs is not yet complete. Mainly the following issues need to be additionally addressed:

1. The force field of an isolated AG must be extended by the most important coupling terms to diminish the frequency deviations. The field dependence of the coupling constants must be determined (which is easy if one assumes linear response).

2. The treatment of the electrostatics in neighboring and covalently linked AGs needs to be scrutinized to check the description of the dipole-dipole coupling for this very special structural motif. For this purpose DFT calculations and DFT/MM simulations on dipeptides will be required as references. Note in this connection that the so-called "correction map (CMAP)", which has been suggested by MacKerell[58,59] and fixes certain CHARMM errors caused by the usual dihedral potentials at the C_α atoms linking neighboring AGs, is available in our MD program EGO. Thus an adjustment of the CMAP to our new AG force field can guarantee that peptide free energy landscapes are correct.

3. The use of static partial charges within an AG should be abandoned in favor of a polarizable electrostatics model which may comprise fluctuating charges, atomic inducible dipoles or both (see, e.g., Refs. 25, 26). Also for an optimal choice of such a model comparisons with DFT and DFT/MM results on dipeptides are necessary.

Clearly all these efforts can benefit a lot if, in addition to the computational reference data from DFT and DFT/MM descriptions, new IR data on model compounds will become available. Even in the context of the present study, which concentrated on the properties of a single AG, the limited amount of available experimental data on the relevant model compound NMA was a matter of concern. In this case only IR data for a single isotopomer (i.e. the N–D derivative) were available, gas phase data were of limited quality and covered a too small spectral range, and so on.

Once all these additional steps will be taken, a computational approach towards the IR spectra of polypeptide backbones will become available, which should be capable of predicting these spectra (at the limited computational cost of a short MM-MD simulation of a peptide/solvent system) more accurately than the previous excitonic TDC models. Because in applications to peptides and proteins the backbone IR spectra will be derived from MD trajectories of the backbone's dipole moment by Fourier transformation, and because MD simulations cover all electrostatic interactions within peptide/solvent simulation systems, the coupling of the amide dipoles is included by construction.

Acknowledgments

This work was supported by the Boehringer Ingelheim Fonds, the Volkswagenstiftung (I 79/884), and by the Deutsche Forschungsgemeinschaft (SFB 533/C3).

Supporting information available

Beyond the above material we have provided seven additional figures and one table together with explanations and captions. This information is available free of charge via the Internet at http://pubs.acs.org.

References

[1] Byler, D. M.; Susi, H. *Biopolymers* **1986,** *25,* 469–487.

[2] Barth, A.; Zscherp, C. *Q. Rev. Biophys.* **2002,** *35,* 369–430.

[3] Siebert, F. *Meth. Enzymol.* **1995,** *246,* 501–526.

[4] Kohn, W.; Sham, L. J. *Phys. Rev.* **1965,** *140,* 1133–1138.

[5] Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964,** *136,* 864–871.

[6] Nonella, M.; Tavan, P. *Chem. Phys.* **1995,** *199,* 19–32.

[7] Neugebauer, J.; Hess, B. A. *J. Chem. Phys.* **2003,** *118,* 7215–7225.

[8] Schmitz, M.; Tavan, P. On the art of computing the IR spectra of molecules in condensed phase. In *Modern methods for theoretical physical chemistry of biopolymers*; Tanaka, S.; Lewis, J., Eds.; Elsevier: Amsterdam, 2006; Chapter 8, pages 157–177.

[9] Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1999,** *110,* 10452–10467.

[10] Bour, P.; Keiderling, T. A. *J. Phys. Chem. B* **2005,** *109,* 5348-5357.

[11] Schrader, T. E.; Schreier, W. J.; Cordes, T.; Koller, F. O.; Babitzki, G.; Denschlag, R.; Renner, C.; Löweneck, M.; Dong, S.-L.; Moroder, L.; Tavan, P.; Zinth, W. *Proc. Natl. Acad. Sci. (USA)* **2007,** *104,* 15729–15734.

[12] Krimm, S.; Bandekar, J. *Adv. Prot. Chem.* **1986,** *38,* 181–364.

[13] Torii, H.; Tasumi, M. *J. Chem. Phys.* **1992,** *97,* 86–91.

[14] Torii, H.; Tasumi, M. *J. Chem. Phys.* **1992,** *97,* 92–98.

[15] Ham, S.; Kim, J. H.; Lee, H.; Cho, M. H. *J. Chem. Phys.* **2003,** *118,* 3491–3498.

[16] Lee, H.; Kim, S. S.; Choi, J. H.; Cho, M. *J. Phys. Chem. B* **2005,** *109,* 5331–5340.

[17] Bour, P.; Keiderling, T. A. *J. Chem. Phys.* **2003,** *119,* 11253–11262.

[18] Torii, H. *J. Phys. Chem. B* **2007,** *111,* 5434–5444.

[19] Zhuang, W.; Abramavicius, D.; Hayashi, T.; Mukamel, S. *J. Phys. Chem. B* **2006,** *110,* 3362–3374.

[20] Kubelka, J.; Keiderling, T. A. *J. Phys. Chem. A* **2001,** *105,* 10922–10928.

[21] Lumley Jones, R. *J. Mol. Spectrosc.* **1963,** *11,* 411–421.

[22] Mayne, L. C.; Hudson, B. *J. Phys. Chem.* **1991,** *95,* 2962–2967.

[23] Chen, X. G.; Schweitzerstenner, R.; Asher, S. A.; Mirkin, N. G.; Krimm, S. *J. Phys. Chem.* **1995,** *99,* 3074–3083.

[24] Ahlstrom, P.; Wallqvist, A.; Engstrom, S.; Jonsson, B. *Mol. Phys.* **1989,** *68,* 563–581.

[25] Palmö, K.; Mannfors, B.; Mirkin, N. G.; Krimm, S. *Biopolymers* **2003,** *68,* 383–394.

[26] Harder, E.; Kim, B. C.; Friesner, R. A.; Berne, B. J. *J. Chem. Theory. Comp.* **2005,** *1,* 169–180.

[27] Guillot, B. *J. Mol. Liq.* **2002,** *101,* 219-260.

[28] Ponder, J. W.; Case, D. A. Force fields for protein simulations. In *Protein Simulations*, Vol. 66; Academic Press Inc: San Diego, 2003.

[29] Tavan, P.; Carstens, H.; Mathias, G. Molecular dynamics simulations of proteins and peptides: Problems, achievements, and perspectives. In *Protein Folding Handbook*; Buchner, J.; Kiefhaber, T., Eds.; Wiley-VCH: Weinheim, 2005.

[30] Schmitz, M.; Tavan, P. *J. Chem. Phys.* **2004,** *121,* 12233–12246.

[31] Schmitz, M.; Tavan, P. *J. Chem. Phys.* **2004,** *121,* 12247–12258.

[32] Becke, A. D. *Phys. Rev. A* **1988,** *38,* 3098–3100.

[33] Perdew, J. P. *Phys. Rev. B* **1986,** *33,* 8822–8824.

[34] Schropp, B.; Tavan, P. *J. Phys. Chem. B* **2008,** *(in press),*.

[35] MacKerell, A. D. *et al. J. Phys. Chem. B* **1998,** *102,* 3586–3616.

[36] Lifson, S.; Stern, P. S. *J. Chem. Phys.* **1982,** *77,* 4542–4550.

[37] Palmö, K.; Pietilä, L.-O.; Krimm, S. *J. Comp. Chem.* **1991,** *12,* 385–390.

[38] Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003,** *118,* 10847–10860.

[39] Boatz, J. A.; Gordon, M. S. *J. Phys. Chem.* **1989,** *93,* 1819–1826 33.

[40] Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989,** *162,* 165.

[41] Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995,** *102,* 346–354.

[42] Hutter, J. et al. *CPMD V3.9.2, Copyright IBM research division, MPI für Festkörperforschung Stuttgart, see www.cpmd.org;* 2005.

[43] Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991,** *43,* 1993–2006.

[44] Nonella, M.; Mathias, G.; Eichinger, M.; Tavan, P. *J. Phys. Chem. B* **2003,** *107,* 316–322.

[45] Nonella, M.; Mathias, G.; Tavan, P. *J. Phys. Chem. A* **2003,** *107,* 8638–8647.

[46] Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994,** *100,* 5829–5835.

[47] Urey, H. C.; Bradley, C. A. *Phys. Rev.* **1931,** *38,* 1969–1978 16.

[48] Singh, U. C.; Kollman, P. A. *J. Comp. Chem.* **1984,** *5,* 129–145.

[49] We used the nonlinear least-squares Marquardt-Levenberg algorithm as implemented in gnuplot, Linux version 3.7, see http://www.ucc.ie/gnuplot.

[50] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983,** *79,* 926–935.

[51] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984,** *81,* 3684–3690.

[52] Niedermeier, C.; Tavan, P. *J. Chem. Phys.* **1994,** *101,* 734–748.

[53] Niedermeier, C.; Tavan, P. *Molecular Simulation* **1996,** *17,* 57–66.

[54] Verlet, L. *Phys. Rev.* **1967,** *159,* 98–103.

[55] Borysow, J.; Moraldi, M.; Frommhold, L. *J. Mol. Phys.* **1985,** *56,* 913–922.

[56] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing;* Cambridge University Press: Cambridge, 1992.

[57] Klähn, M.; Mathias, G.; Kötting, C.; Nonella, M.; Schlitter, J.; Gerwert, K.; Tavan, P. *J. Phys. Chem. A* **2004,** *108,* 6186–6194.

[58] MacKerell, A. *J. Comp. Chem.* **2004,** *25,* 1584–1604.

[59] MacKerell, A.; Feig, M.; Brooks, C. *J. Comp. Chem.* **2004,** *25,* 1400–1415.

Supporting Information to the article: A polarizable force field for computing the infra-red spectra of the polypeptide backbone

Verena Schultheis, Rudolf Reichold, Bernhard Schropp, and Paul Tavan*

Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität,

Oettingenstr. 67, 80538 München, Germany

*email: tavan@physik.uni-muenchen.de; phone: +49-89-2180-9220

The sets of graphs in Figs. 12 and 13 document the results of the MT/BP calculations on NMA exposed to homogeneous electric fields E_j oriented along the axes $j \in \{x, y, z\}$ of the Cartesian coordinate system depicted in Fig. 2. Field dependent equilibrium values of the internal coordinates depicted in Fig. 3 were taken directly from NMA's equlibrium geometries $\mathcal{G}(E_j)$, whereas harmonic force constants $k_i(E_j)$ were derived from the MT/BP Hessians $\mathbf{H}(E_j)$ through the iterative match of intrinsic frequencies explained in Sec. 2.2. The graphs contain linear fits to the MT/BP derived force field parameters whose slopes give the linear response parameters in Tab. 4. Some of the deviations from linearity occurring at very large fields are computational artifacts of the MT/BP approach associated with a field-induced pushing of electron density towards the boundaries of the DFT box, whithin which that density is expanded into plane waves.

Fig. 14 serves to illustrate the difference between the simple AG force field defined by Eqs. (2) and (4) and a DFT force field. By visually subtracting the nearly diagonal MM(DFT) Hessian in (B) from the DFT reference in (A) one can get clues, which of the neglected coupling terms given by Eq. (6) are important for the computation of correct frequencies.

Fig. 15 demonstrates that most parts of the IR spectra of protonated and deuterated NMA in aqueous solution, which were derived by FTTCF from extended DFT/MM-MD simulations, are quite well converged. Mainly the regions of the AI bands exhibit larger differences in spectral shape when computed from the first 26 ps (red) and all 78 ps (blue) of the simulations. We suspect that in both cases the respective system was not yet sufficiently equilibrated at the beginning of the production run.

Fig. 16 shows for the case of protonated NMA that the effects of hydrogen bonding on solvatochromic band shifts are small. See the caption to this figure and the discussion of the dashed line in Fig. 11D.

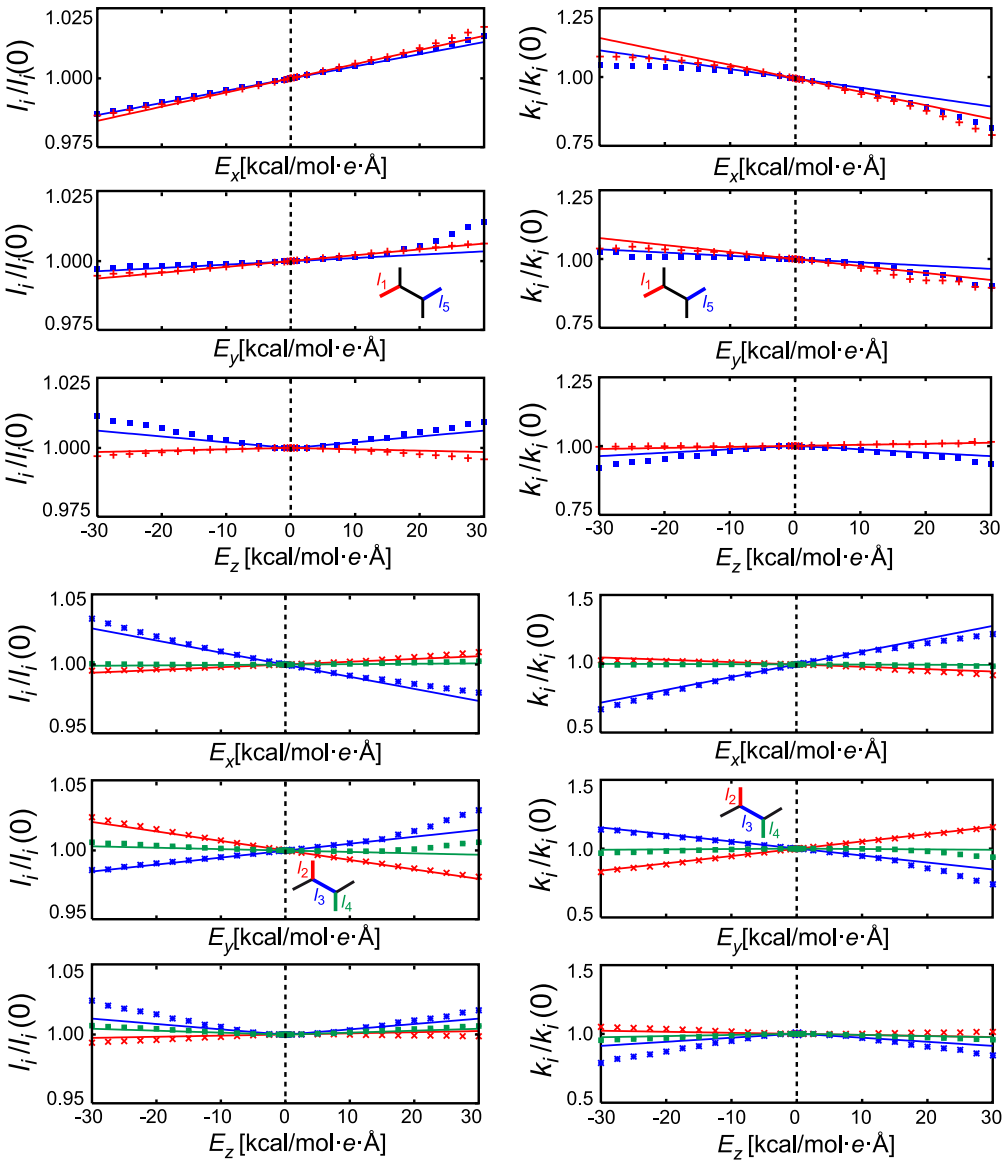

Figure 12: MT/BP equilibrium bond lengths $l_i(\mathbf{E})/l_i(0)$ (left) and force constants $k_i(\mathbf{E})/k_i(0)$ (right) as functions of electric fields E_x, E_y, E_z; solid lines are linear fits to data for fields in the range $[-10, 10]$ kcal/(mol Å e).

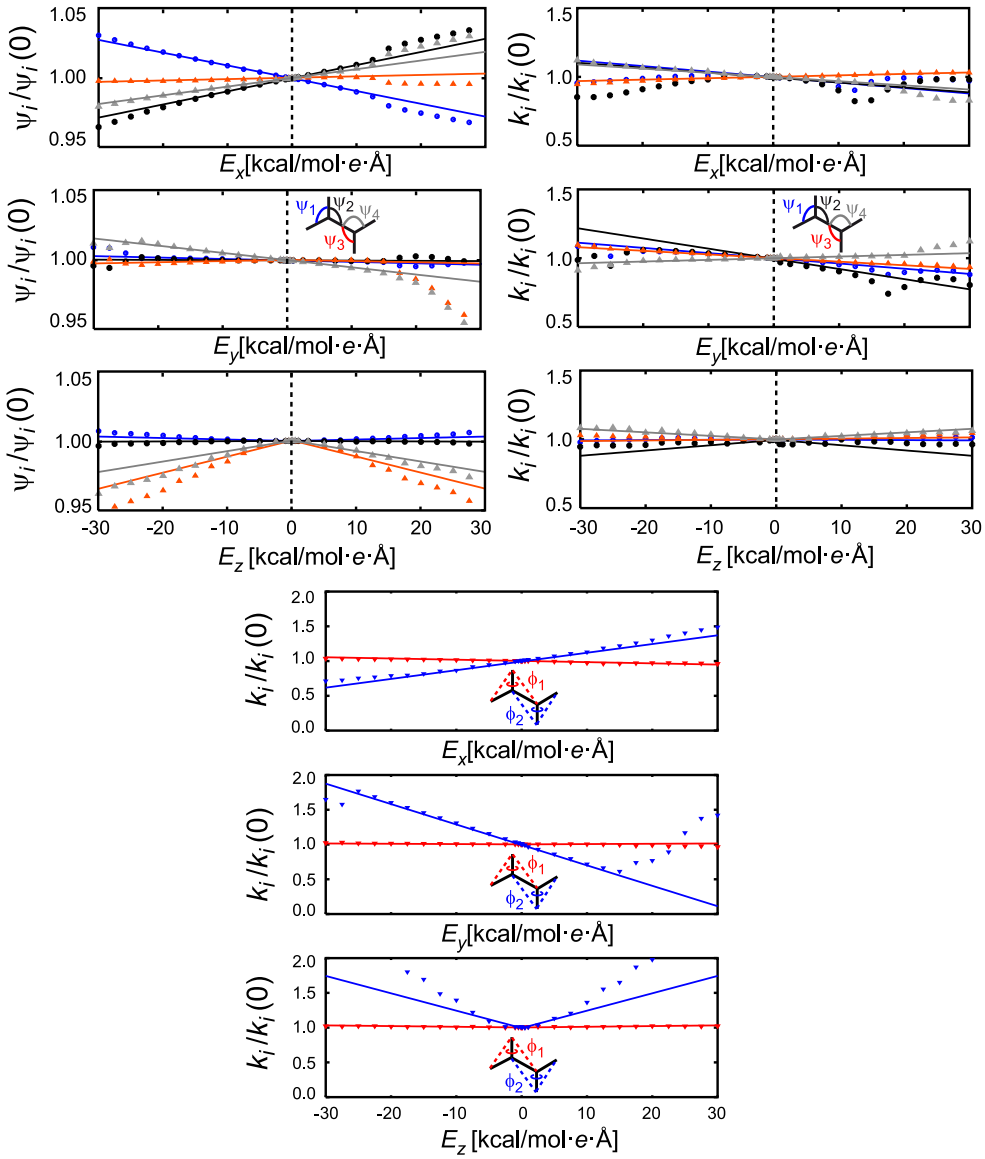

Figure 13: MT/BP equilibrium bond angles $\psi_i(\mathbf{E})/\psi_i(0)$ (top left), associated force constants $k_i(\mathbf{E})/k_i(0)$ (top right) and force constants of improper dihedral angles $\phi_i(\mathbf{E})/\phi_i(0)$ as functions of electric fields E_x, E_y, E_z and associated linear fits.

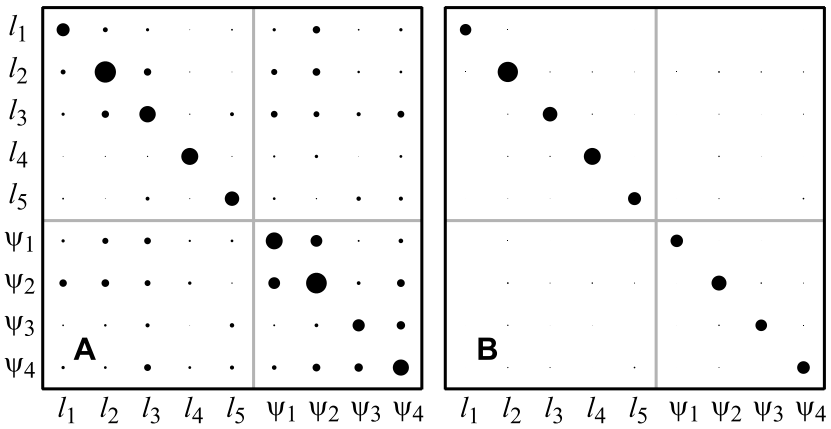

Figure 14: Structure of Hessian matrices in internal coordinates from (A) DFT and (B) MM(DFT) force fields for isolated NMA molecules. Absolute values of matrix elements are coded by circle sizes. The small off-diagonal terms in (B) are due to small so-called non-bonded interactions within NMA.

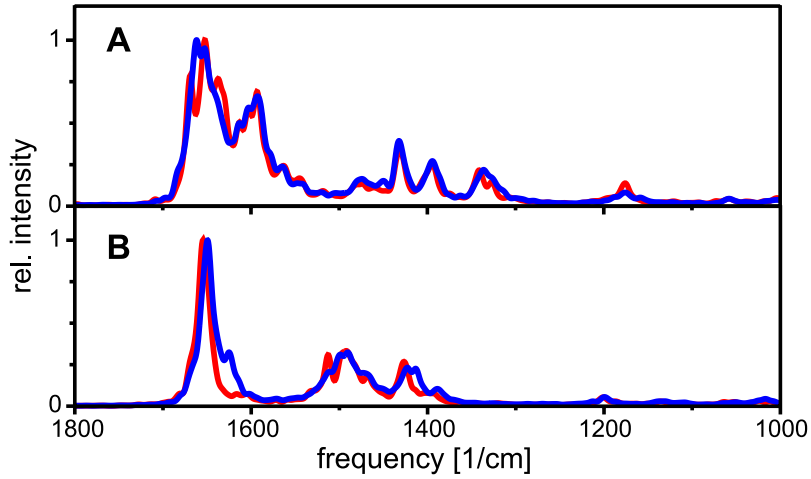

Figure 15: Convergence of solution spectra for (A) protonated and (B) deuterated NMA. FTTCF spectra from DFT/MM-MD trajectories of different lengths are compared: First 26 ps (red), and all 78 ps (blue). Frequencies are scaled by a factor of 1.0354. Intensities are normalized to the respective maximum values.

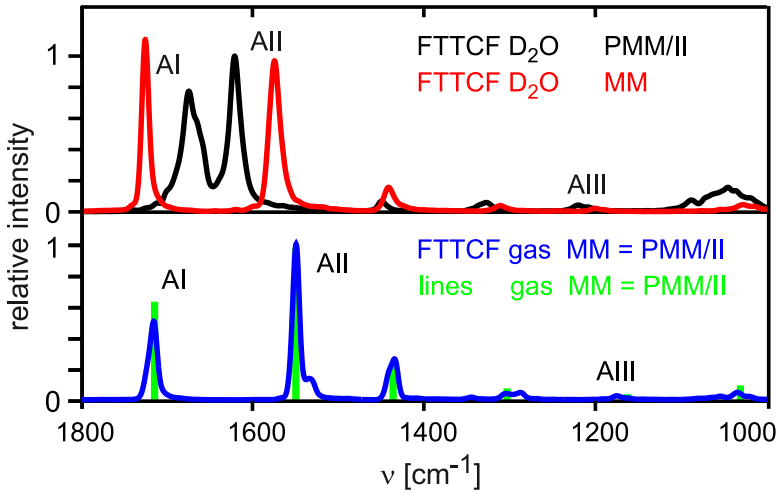

Figure 16: Comparison of FTTCF spectra calculated for protonated NMA in D_2O and in the gas phase (bottom) with different MM force fields from MD trajectories. By comparing the MM result for D_2O (red) with that for the gas phase (blue) one gets a clue on the effects of hydrogen bonding, by comparing with the PMM/II result for D_2O (black) one recognizes the effects of bond polarization. For the gas phase the MM(DFT) line spectrum obtained from normal mode analysis is also given. Frequencies are scaled by a factor of 0.9925.

The remaining Figs. 17, 18, and Tab. 5 document a failed attempt to construct a PMM/II force field for AGs, which performs better at least on the frequencies of the AI-AIII modes while retaining the quality of the mode description achieved with the force field MM(DFT). In this attempt we considered the force constants of the ICs as adjustable parameters and tried to optimize by gradient descent the deviation measure Eq. (10) defined further below (using the index sets $\mathcal{I} = \{1, \ldots, 6\}$ and $\mathcal{J} = \{7, 8, 9\}$).

The modified force constants obtained by these gradient descent procedures are listed in Tab. 5 under the labels $\mathrm{MM_{o1}}$(DFT) and $\mathrm{MM_{o2}}$(DFT) together with the starting values MM(DFT). By construction the optimized force fields closely reproduce the MT/BP spectrum of the "AG within NMA" in the high frequency region as can be seen by comparing the corresponding columns in Fig. 17. On the other hand, according to Fig. 18 the frequency-optimized force fields $\mathrm{MM_{o1/2}}$(DFT) now predict for the normal modes AI-AIII compositions, which sizably deviate from the MT/BP references. Thus by a tuning of the harmonic force constants one can reach an excellent match of selected frequencies while loosing the perfect match of the corresponding normal mode compositions. As a result the inclusion of the cross terms (5) cannot be avoided for an improved frequency matching that retains good quality mode compositions.

In our frequency optimizations we used the weighted deviation measure

$$\langle \Delta\nu^2 \rangle_{\mathcal{I}}^{\mathcal{J}} = \sqrt{\frac{\sum_{i\in\mathcal{I}} \Delta\nu_i^2 + (\sum_{i\in\mathcal{J}} \Delta\nu_i^2)/20}{|\mathcal{I}| + |\mathcal{J}|/20}} \tag{10}$$

where $\mathcal{I}$ and $\mathcal{J}$ are two sets of indices. The frequencies ν_i, $i \in \mathcal{I}$, fully contribute to this deviation whereas frequencies ν_i, $i =\in \mathcal{J}$ only contribute with a weight reduced by a factor 1/20. This measure takes all frequencies into account but is mainly determined by frequencies with indices $i \in \mathcal{I}$.

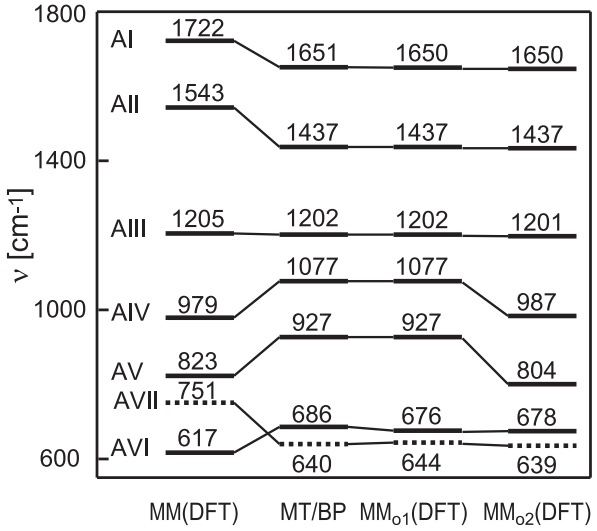

Figure 17: MT/BP spectra of the "AG within NMA" (defined by methyl groups with nearly massless hydrogen atoms, cf. Sec. 4.3) compared with MM spectra calculated at different parameterizations. The employed parameter sets MM(DFT), MM_{o1}(DFT), and MM_{o2}(DFT) are listed in the corresponding columns of Tab. 5 and are explained in the text. For line styles and mode labels see the caption to Fig. 4.

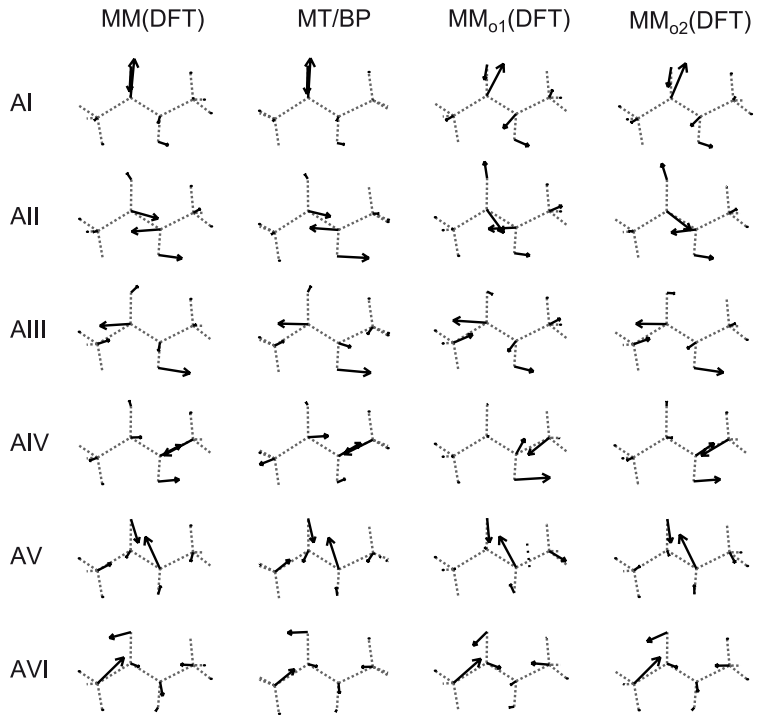

Figure 18: In-plane normal modes of the "AG within NMA" (cf. Sec. 4.3) belonging to the line spectra depicted in Fig. 17. Dashed lines indicate the bonds within the NMA molecule covering the AG depicted in Fig. 3. Arrows indicate atomic motions.

Table 5: Harmonic zero-field parameters of an AG.

		MM(DFT)	MM_{o1}(DFT)	MM_{o2}(DFT)
q_i	q_i^0	k_i	k_i	k_i
l_1	1.520 Å	460*	629*	577*
l_2	1.233 Å	1443*	1075*	1214*
l_3	1.374 Å	748*	543*	490*
l_4	1.015 Å	1009*	1004*	1006*
l_5	1.455 Å	591*	781*	604*
ψ_1	121.9 deg	4.7$	4.2$	5.6$
ψ_2	122.6 deg	6.9$	5.1$	7.4$
ψ_3	118.5 deg	4.2$	3.8$	3.9$
ψ_4	122.5 deg	4.9$	14.2$	7.3$
ϕ_1	0 deg	7.1$	4.9$	4.9$
ϕ_2	0 deg	1.2$	1.1$	1.1$
ξ	180 deg	0.8$	0.8$	0.8$

*[kcal/(mol Å^2)], $[kcal/{mol (10 deg)^2}].

4 Zusammenfassung und Ausblick

Proteine weisen neben schnellen, thermischen Fluktuationen auch langsame Übergänge zwischen einigen wenigen metastabilen Zuständen, den sogenannten Konformationen, auf. Diese Konformationsdynamik ist entscheidend für die biologische Funktion der Proteine. Der Abdruck des Artikels „Extracting Markov models of peptide conformational dynamics from simulation data“ [24] in Kapitel 2 erläutert statistische Verfahren zur Zeitreihenanalyse, welche es erlauben, Trajektorien dieser komplexen hochdimensionalen Systeme zu untersuchen und die metastabilen Zustände der zugrunde liegenden Dynamik anhand ihrer Lebensdauer hierarchisch zu klassifizieren.

Dazu wurde das Modell eines neuronalen Netzes [115, 116], das ursprünglich als Clustering-Verfahren entwickelt wurde, so modifiziert, dass es eine dichteorientierte Diskretisierung für den Datenraum verwendet, welche Verzerrungen der Metrik des Datenraumes vermeidet. Die für diese Diskretisierung verwendete unscharfe Partitionierung des Datenraumes mit Hilfe einer Mischung aus R univariaten Normalverteilungen, deren Zentren das sogenannte Kodebuch bilden, teilt jeden Punkt der Trajektorie mehreren Normalverteilungen mit unterschiedlichen Zuordnungswahrscheinlichkeiten $\mathbf{p}$ zu. Die Partitionierung dient auch zur Bestimmung einer R-dimensionalen Übergangsmatrix, die die Dynamik des analysierten Systems als Markovmodell beschreibt.

Zur Analyse dieser Übergangsmatrix dient eine nichtlineare Dynamik im Raum der Zuordnungswahrscheinlichkeiten $\mathbf{p}$, die eine Mischung ist aus einem linearen Diffusionsanteil, der einen reinen Markov-Prozess beschreibt, und einem nichtlinearen Term, der eine Darwinsche Selektion beinhaltet, die dem Diffusionsterm entgegenwirkt. Das Verhältnis dieser gegensätzlichen Prozesse wird von einem Gewichtungsparameter κ gesteuert. Für $\kappa = 0$ ergibt sich eine rein lineare Wahrscheinlichkeitsdiffusion, die bei einem einfach zusammenhängenden Datenraum unabhängig vom Startwert zu derselben stationären Verteilung der Zuordnungswahrscheinlichkeiten $\mathbf{p}$ führt. Mit zunehmenden κ steigt die Anzahl der stationären Punkte der nichtlinearen Wahrscheinlichkeitsdynamik. Als Prototypen der Dynamik, welche der Trajektorie zugrunde liegt, kennzeichnen diese stationären Punkte metastabile Zustände, die gesuchten Konformationen. Die virtuelle Dichte bildet die Wahrscheinlichkeitsdichten $\mathbf{p}$ auf

den Datenraum ab. Dort sind die Konformationen die ersten Momente der virtuellen Dichte, während die Varianz der virtuellen Dichte die Ausdehnung der Konformationen im Datenraum misst.

Die Anzahl der Prototypen wächst monoton und ist über weite Bereiche von κ konstant. Die Größe dieser konstanten Bereiche ist ein Maß für die Plausibilität der betreffenden Modelle. Innerhalb eines solchen Bereichs ändert sich die Lage der Prototypen kaum, wenn man von den Werten kurz nach einer Aufspaltung absieht. Daher genügt es, sich auf diejenigen Werte von κ zu beschränken, bei denen gerade noch keine Abspaltung weiterer Prototypen auftritt. Für diese Werte von κ kann man mit Hilfe der nichtlinearen Dynamik eine hierarchische Beziehung zwischen den Prototypen für je zwei Werte von κ aufstellen. Außerdem erlaubt es die nichtlineare Dynamik, den gesamten Datensatz auf jeder Hierarchieebene zu klassifizieren, also jeden Punkt der Trajektorie einem Prototypen zuzuordnen. Auch das Kodebuch kann so klassifiziert werden.

Auf jeder Ebene dieser Hierarchie von Prototypen lässt sich die zugehörige Markov-Matrix auf verschiedene Weisen bestimmen: Dies erfolgt durch Reduktion der ursprünglichen R-dimensionalen Übergangsmatrix unter Verwendung der Klassifikation des Kodebuchs, direkt aus der Klassifikation der Daten oder durch Abzählen der Übergänge zwischen klassifizierten Daten. Die Größe der Diagonalelemente der daraus resultierenden Markov-Matrizen ergibt die Lebensdauern der betreffenden Zustände.

Alternativ zu diesem Vorgehen kann man die Hierarchie aus Prototypen auch dadurch aufstellen, dass man, von der ursprünglichen R-dimensionalen Markov-Matrix ausgehend, iterativ diejenigen Partitionsvolumina vereinigt, zwischen denen die schnellsten Übergänge erfolgen. Dieses Verfahren basiert auf dem Prinzip der in Abschnitt 1.5.2 angesprochenen detaillierten Bilanz. Auch diese Methode liefert ein Maß für die Zeitskala, auf der die schnellsten Übergänge jedes Modells der Hierarchie stattfinden. Anhand dieser Zeitskalen kann man ein plausibles Modell zur Beschreibung der Dynamik auswählen.

Die Funktion aller diskutierten Algorithmen wurde, außer an einem konstruierten eindimensionalen Beispieldatensatz, durch die Analyse einer Trajektorie eines Tripeptids demonstriert. Alle Varianten des beschriebenen Verfahrens liefern vergleichbare Ergebnisse. Weiterhin illustriert Kapitel 2, welche Auswertungsmöglichkeiten die Methode bietet: Aus den relativen Häufigkeiten der Zustände lassen sich Unterschiede ihrer freien Energien bestimmen. Außerdem liefert die graphische Darstellung der reduzierten Matrizen eine Visualisierung der Konnektivität der einzelnen Zustände.

Zusätzlich können farbkodierte Abbildungen der Trajektorie die Klassifikation der Trajektorienpunkte visualisieren.

Prinzipiell lassen sich die beschriebenen Verfahren auch auf andere Systeme anwenden, bei denen es darum geht, die Charakteristika langsamer Übergänge zwischen wenigen metastabilen Zustände von schnellen Fluktuationen zu trennen. Als Beispiel hierfür dient die Analyse der Konformationsdynamik eines Disaccharids und eines Trisaccharids, die ich mit den hier beschriebenen Methoden in Zusammenarbeit mit Austin B. Yongye, Jorge Gonzalez-Outeiriño, John Glushka und Robert J. Woods vom Complex Carbohydrate Research Center der University of Georgia durchgeführt habe [125].

Die beschriebenen Verfahren erlauben es, Informationen über die Konformationsdynamik biologischer Systeme aus simulierten Trajektorien zu gewinnen. Ähnliche Fragestellungen werden experimentell mit Hilfe der Fourier-Transformations-Infrarot-(FTIR-)Spektroskopie untersucht. Experimentatoren verwenden häufig empirische Regeln, um aus IR-Spektren Rückschlüsse über den Gehalt an Sekundärstrukturmotiven, wie α-Helices und β-Faltblätter, zu gewinnen. Prinzipiell lassen sich die IR-Spektren mit rechenaufwändigen quantenmechanischen Simulationsmethoden berechnen. Praktisch übersteigt der Aufwand dafür jedoch bereits für verhältnismäßig kleine Peptide die Möglichkeiten modernster Rechencluster, insbesondere wenn diese Moleküle in ihrer biologischen Umgebung, das heißt in Wasser, simuliert werden sollen.

Obwohl klassische Molekülmechanik-(MM-)Simulationen mit herkömmlichen Kraftfeldern sonst häufig als weniger aufwändige Alternative zu quantenmechanischen Simulationen verwendet werden, sind sie für derartige Rechnungen wenig geeignet, weil sie die Polarisation der Proteine durch die Lösungsmittelumgebung vernachlässigen. Die Polarisation führt jedoch zu großen Verschiebungen der IR-Frequenzen. Neuere, sogenannte polarisierbare Kraftfelder berücksichtigen zwar den Einfluss der Polarisation in der Elektrostatikberechnung, vernachlässigen die Polarisationseffekte aber nach wie vor für die Potentiale, die die Steifheit der chemischen Bindungen bezüglich Streckungen, Winkelverbiegungen und Torsionen modellieren.

Kapitel 3 erläutert ein aus dieser Kritik hervorgegangenes Kraftfeld, das speziell für die Berechnung von Spektren im mittleren Infrarotbereich entwickelt wurde. Dieses Kapitel enthält dazu ein Manuskript [124], das ich gemeinsam mit Rudolf Reichold, Bernhard Schropp und Paul Tavan zur Veröffentlichung eingereicht habe, sowie ergänzende Grafiken.

Der Vergleich zweier mit unterschiedlichen Dichtefunktionaltheorie-(DFT-)Methoden berechneter Vakuum-Spektren untereinander und mit experimentellen Ergebnissen zeigt, dass DFT-Rechnungen grundsätzlich als Vorlage zur Entwicklung eines Kraftfelds für IR-Spektren des Proteinrückgrats geeignet sind, wenn man die Frequenzen der Moden geeignet skaliert.

Zur Parametrisierung des Kraftfelds dienten daher DFT-Simulationen und experimentelle Messungen am Molekül N-Methylacetamid (NMA, vgl. Abbildung 1.2), das alle Atom- und Bindungstypen eines Peptidplättchens enthält, also derjenigen sechsatomigen Einheit, aus der das Proteinrückgrat aufgebaut ist. Die Parameter des Kraftfelds (vgl. Abbildung 3 in Kapitel 3) umfassen alle fünf Bindungslängen innerhalb des Peptidplättchens, einen nicht-redundanten Satz aus vier Bindungswinkeln, zwei *improper* und einen echten Diederwinkel. Die übrigen Freiheitsgrade des Moleküls NMA, die die Methylgruppen außerhalb des Peptidplättchens beschreiben, spielen für die Beschreibung des Proteinrückgrats keine Rolle und wurden daher nicht optimiert.

Das Kraftfeld verwendet, wie beispielsweise beim CHARMM-Kraftfeld [80] üblich, harmonische Potentiale für Bindungslängen, Bindungswinkel und Torsionswinkel zur Beschreibung kurzreichweitiger Wechselwirkungen zwischen eng benachbarten Atomen. Deren Ruhelagen lassen sich über quantenmechanische Geometrieoptimierungen beispielsweise mit der DFT verhältnismäßig einfach bestimmen.

Um die Kraftkonstanten der harmonischen Potentiale zu berechnen, bestimmt man, ausgehend von der schon berechneten optimalen Geometrie, die Hesse-Matrix in DFT-Rechnungen durch finite Differenzen. Die Methode von Boatz und Gordon [126] leitet aus der Hesse-Matrix eine intrinsische Frequenz für jeden Freiheitsgrad des Moleküls ab. Iterative Variation der Parameter eines MM-Kraftfelds für diese Freiheitsgrade bis zum Erreichen der gleichen intrinsischen Frequenzen lieferte die gesuchten Kraftkonstanten. Um dabei den Einfluss der Methylgruppen des NMA-Moleküls auszuschalten, wurden die Massen der Methyl-Wasserstoffatome soweit reduziert, dass die zugehörigen Methylschwingungen mindestens doppelt so hohe Frequenzen hatten wie die höchste Amidfrequenz, die NH-Streckschwingung. Dadurch sind Kopplungen zwischen Methylschwingungen und Amidschwingungen ausgeschlossen. Außerdem wurden die Methyl-Wasserstoffatome entladen (wobei die Ladung der Methyl-Kohlenstoffatome so gewählt wurde, dass das NMA Molekül dennoch neutral blieb) und ihre van-der-Waals-Wechselwirkung vernachlässigt. Alle diese Maßnahmen dienen dazu, den Einfluss der Methyl-Parametrisierung auf die Parameter des Peptidplättchens möglichst gering zu halten.

Die mit diesem Parametersatz MM(DFT) berechneten Spektren weichen in den Frequenzen der Amidbanden von entsprechenden DFT-Rechnungen ab. Die Modenzusammensetzungen entsprechen jedoch sehr genau denen der DFT-Rechnungen. Eine weitere Parametrisierung MM_{o1}(DFT) erfolgte anhand eines speziellen Optimierungskriteriums, das die gewichteteten mittleren Abweichungen zwischen den DFT- und den zu optimierenden MM-Amidfrequenzen berücksichtigt. Dadurch ergeben sich MM_{o1}(DFT)-Amidfrequenzen, die konstruktionsgemäß sehr genau mit denen der DFT-Rechnung übereinstimmen. Allerdings zeigen sich hier deutliche Veränderungen der Modenzusammensetzungen. Dies war der Anlass für einen weiteren Optimierungsversuch MM_{o2}(DFT), der sich auf die Optimierung der Amid-I bis III Moden konzentrierte, und dadurch die Modenzusammensetzung weniger beeinflusste. Insgesamt zeigen diese Optimierungsversuche, dass ein Kompromiss zwischen exakten Frequenzen und realistischen Modenzusammensetzungen prinzipiell möglich ist. Analog zur Optimierung MM_{o1}(DFT) erfolgte zusätzlich die Optimierung MM_{o1}("exp") anhand eines mit den bereits angesprochenen modenspezifischen Skalierungsfaktoren für die DFT-Rechnung konstruierten „experimentellen“ Spektrums.

Um die Feldabhängigkeit der Kraftfeldparameter zu bestimmen, erfolgte eine Reihe von DFT-Rechnungen für NMA in homogenen externen elektrischen Feldern. Es zeigte sich, dass die Abhängigkeit sowohl der Ruhelagen als auch der Kraftkonstanten gut linear beschreibbar ist.

Als Vergleich enthält Referenz [124] ein IR-Spektrum, das mittels FTTCF (vgl. Abschnitt 1.3) aus einer ausgedehnten DFT/MM-Simulation von NMA in Wasser berechnet wurde. Dieses Spektrum stimmt sehr gut mit experimentellen Messungen überein, was noch einmal zeigt, dass die verwendeten Methoden prinzipiell geeignet sind zur Simulation der Trajektorie und zur Berechnung der Spektren. Außerdem zeigen die aus MM-Rechnungen mit dem neuen Kraftfeld berechneten Solavatochromieeffekte gute Übereinstimmung mit denen aus DFT-Rechnungen. Damit ist der Beweis des Prinzips gelungen.

Für die weitere Verbesserung der Verfahren bieten sich folgende Ansätze: Bisher ist der Einfluss der Polarisationseffekte auf die Ladungsverteilung innerhalb des Peptidplättchens unberücksichtigt geblieben. Dynamische, feldabhängige Ladungsverschiebungen, wie sie in Abschnitt 1.4.2 im Zusammenhang mit polarisierbaren Kraftfeldern angesprochen wurden, sollten zu einer realistischeren Ladungsverteilung im Peptidplättchen führen, welche direkt über das Dipolmoment die IR-Spektren beeinflusst. Es bietet sich an, diese Effekte bei der Simulation zu berücksichtigen – ähnlich wie in Referenzen [86, 87] demonstriert. Weiterhin sollte an Stelle der intrinsischen Frequenzen die gesamte Hesse-Matrix in die Optimierung der Kraftkonstanten ein-

gehen, um so Kopplungen zwischen einzelnen Freiheitsgraden zu berücksichtigen, welche sich in den Nebendiagonalelementen der Hesse-Matrix zeigen. Auch der Einfluss der Proteinseitenketten, das heißt der Aminosäurereste (vgl. Abschnitt 1.1) kann in zukünftigen Weiterentwicklungen berücksichtigt werden.

Neben der Kraftfeldentwicklung erfolgte im Rahmen dieser Arbeit die Implementierung geeigneter Datenstrukturen und Algorithmen in das MM-Programm der Arbeitsgruppe [76], die Simulationen von Proteinen mit dem neuen Kraftfeld auf Einprozessorrechnern oder auf parallelen Rechenclustern erlauben. Erste Vergleiche an NMA in Wasser ergaben, dass diese klassische Berechnung über 400 mal schneller ist als entsprechende DFT/MM-Rechnungen. Bei größeren Molekülen dürfte sich eine noch größere Rechenzeitersparnis ergeben. Damit rücken viele Berechnungen, die bisher am Rechenaufwand gescheitert sind, in greifbare Nähe: Sowohl die Berechnung von größeren Peptiden oder Proteinen erscheint möglich, als auch die Berechnung zeitaufgelöster Spektren oder die Berücksichtigung größerer Ensembles von Strukturen, die dann statistisch besser abgesicherte Aussagen erlauben.

Literaturverzeichnis

[1] Oesterhelt, D. und W. Stoeckenius. Functions of a new photoreceptor membrane. *Proc. Natl. Acad. Sci. USA* **70**, 2853–2857 (1973).

[2] Creighton, T. E. *Proteins: structures and molecular properties*. W. H. Freeman and Company, New York, 2. Auflage (1993).

[3] Alzheimer, A. Über eine eigenartige Erkrankung der Hirnrinde. *Allg. Z. Psychiat. Psych.-Gerichtl. Med.* **1–2**, 146–148 (1907).

[4] Creutzfeldt, H. G. Über eine eigenartige herdförmige Erkrankung des Zentralnervensystems. *Z. Gesamte Neurol. Psychiatrie* **57**, 1–18 (1920).

[5] Jakob, A. Über eigenartige Erkrankungen des Zentralnervensystems mit bemerkenswertem anatomischem Befunde (spastische Pseudosklerose-Encephalomyelopathie mit disseminierten Degenerationsherden). *Z. Gesamte Neurol. Psychiatrie* **64**, 147–228 (1921).

[6] Chen, S., F. A. Ferrone und R. Wetzel. Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation. *Proc. Natl. Acad. Sci. USA* **99**, 11884–11889 (2002).

[7] Parkinson, J. *An Essay on the Shaking Palsy*. Sherwood, Neely, and Jones, London (1817).

[8] Caughey, B. und P. T. Lansbury, Jr. Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. *Annu. Rev. Neurosci.* **26**, 267–298 (2003).

[9] Kirkitadze, M. D., G. Bitan und D. B. Teplow. Paradigm shifts in Alzheimer's disease and other neurodegenerative disorders: the emerging role of oligomeric assemblies. *J. Neurosci. Res.* **69**, 567–577 (2002).

[10] Cohen, F. Protein misfolding and prion diseases. *J. Mol. Biol.* **293**, 313–320 (1999).

[11] Dobson, C. Protein misfolding, evolution and disease. *Trends. Biochem. Sci.* **24**, 329–332 (1999).

[12] Stryer, L. *Biochemie*. Spektrum, Heidelberg (1999).

[13] Schmitz, M. *Entwicklung, Anwendung und Vergleich von Methoden zur Berechnung von Infrarotspektren einzelner Moleküle in polaren Lösungsmitteln*. Dissertation, Ludwig-Maximilians-Universität, München (2004).

[14] Branden, C. und J. Tooze. *Introduction to protein strucure*. Garland Publishing, New York (1999).

[15] Lumley Jones, R. The infrared spectra of some simple N-substituted amides in the vapor state. *J. Mol. Spectrosc.* **11**, 411–421 (1963).

[16] Mirkin, N. G. und S. Krimm. Conformers of trans-N-methylacetamide - Ab initio study of geometries and vibrational spectra. *J. Mol. Struct.* **242**, 143–160 (1991).

[17] Mirkin, N. G. und S. Krimm. Structure of trans-N-methylacetamide - Planar or nonplanar symmetry. *Theochem-J. Mol. Struct.* **334**, 1–6 (1995).

[18] Ham, S., J. H. Kim, H. Lee und M. H. Cho. Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA-nD(2)O complexes. *J. Chem. Phys.* **118**, 3491–3498 (2003).

[19] Bour, P. und T. A. Keiderling. Empirical modeling of the peptide amide I band IR intensity in water solution. *J. Chem. Phys.* **119**, 11253–11262 (2003).

[20] Zhuang, W., D. Abramavicius, T. Hayashi und S. Mukamel. Simulation protocols for coherent femtosecond vibrational spectra of peptides. *J. Phys. Chem. B* **110**, 3362–3374 (2006).

[21] Edler, J. und P. Hamm. Spectral response of crystalline acetanilide and N-methylacetamide: Vibrational self-trapping in hydrogen-bonded crystals. *Phys. Rev. B* **69**, 214301:1–8 (2004).

[22] Hamm, P., M. H. Lim und R. M. Hochstrasser. Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. *J. Phys. Chem. B* **102**, 6123–6138 (1998).

[23] Hunt, N. T. und K. Wynne. The effect of temperature and solvation on the ultrafast dynamics of N-methylacetamide. *Chem. Phys. Lett.* **431**, 155–159 (2006).

[24] Schultheis, V., T. Hirschberger, H. Carstens und P. Tavan. Extracting Markov models of peptide conformational dynamics from simulation data. *J. Chem. Theory Comput.* **1**, 515–526 (2005).

[25] Ramachandran, G. N., C. Ramakrishnan und V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).

[26] Levinthal, C. Molecular dynamics simulations of proteins and peptides: Problems, achievements, and perspectives. In J. T. P. DeBrunner und E. Munck (Herausgeber), *Mossbauer spectroscopy in biological Systems, proceedings of a meeting held at Allerton House, Monticello, Il.*, Seiten 22–24. University of Illinois Press, Urbana (1969).

[27] Frauenfelder, H., S. G. Sligar und P. G. Wolynes. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).

[28] Chapagain, P. P., J. L. Parra, B. S. Gerstman und Y. Liu. Sampling of states for estimating the folding funnel entropy end energy landscape of a model alpha-helical hairpin peptide. *J. Chem. Phys.* **127**, 075103:1–7 (2007).

[29] Blundell, T. L. und L. N. Johnson. *Protein crystallography*. Academic Press, London (1976).

[30] Giacovazzo, C.(Herausgeber), *Fundamentals of crystallogrophy*. Oxford University Press, New York (1992).

[31] Rule, G. S. und T. K. Hitchens. *Fundamentals of protein NMR spectroscopy*. Springer, Dordrecht (2006).

[32] McPherson, A. Current approaches to macromolecular crystallization. *Eur. J. Biochem.* **189**, 1–23 (1990).

[33] Blagova, E. V. und I. P. Kuranova. Crystallization and preparation of protein crystals for X-ray diffraction analysis. *Crystallography Reports* **44**, 513–531 (1999).

[34] Havel, H. *Spectroscopic methods for determining protein structure in solution*. VCH Publishers, Inc., New York (1996).

[35] Hamm, P., M. Lim, W. DeGrado und R. Hochstrasser. Functions of a new photoreceptor membrane. *Proc. Natl. Acad. Sci. USA* **96**, 2036–2041 (1999).

[36] Susi, H. Infrared Spectroscopy – Conformation. *Meth. Enzymol.* **26**, 455–472 (1972).

[37] Günzler, H. und H. Heise. *IR-Spektroskopie: Eine Einführung*. VCH Verlagsgesellschaft mbH, Weinheim (1996).

[38] Haris, P. I. und D. Chapman. Does Fourier-transform infrared spectroscopy provide useful information on protein structures? *Trends Biochem. Sci.* **17**, 328–333 (1992).

[39] Watson, T. M. und J. D. Hirst. Influence of electrostatic environment on the vibrational frequencies of proteins. *J. Phys. Chem. A* **107**, 6843–6849 (2003).

[40] Siebert, F. Infrared-spectroscopy applied to biochemical and biological problems. *Meth. Enzymol.* **246**, 501–526 (1995).

[41] Kubelka, J. und T. A. Keiderling. Ab initio calculation of amide carbonyl stretch vibrational frequencies in solution with modified basis sets. 1. N-methyl acetamide. *J. Phys. Chem. A* **105**, 10922–10928 (2001).

[42] Mayne, L. C. und B. Hudson. Resonance Raman-spectroscopy of N-methylacetamide - overtones and combinations of the C-N stretch (amide II') and effect of solvation on the C=O stretch (amide I) intensity. *J. Phys. Chem.* **95**, 2962–2967 (1991).

[43] Chen, X. G., R. Schweitzer-Stenner, S. A. Asher, N. G. Mirkin und S. Krimm. Vibrational assignments of trans-N-methylacetamide and some of its deuterated isotopomers from band decomposition of IR, visible and resonance Raman-spectra. *J. Phys. Chem.* **99**, 3074–3083 (1995).

[44] Schrader, B.(Herausgeber), *Infra-red and Raman spectroscopy: methods and applications*. VCH Verlagsgesellschaft mbH, Weinheim (1995).

[45] Becker, O. M., A. D. MacKerell Jr., B. Roux und M. Watanabe. *Computational biochemistry and biophysics*. Marcel Dekker, Inc., New York (2001).

[46] Press, W. H., S. A. Teukolsky, W. T. Vetterling und B. P. Flannery. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press (1992).

[47] Nonella, M., G. Mathias, M. Eichinger und P. Tavan. Structures and vibrational frequencies of the quinones in Rb. Sphaeroides derived by a combined density functional / molecular mechanics approach. *J. Phys. Chem. B* **107**, 316–322 (2003).

[48] Klähn, M., G. Mathias, C. Kötting, M. Nonella, J. Schlitter, K. Gerwert und P. Tavan. IR Spectra of phosphate ions in aqueous solution: Predictions of a DFT/MM approach compared with observations. *J. Phys. Chem. A* **108**, 6186–6194 (2004).

[49] Schmitz, M. und P. Tavan. Vibrational spectra from atomic fluctuations in dynamics simulations. I. Theory, limitations, and a sample application. *J. Chem. Phys.* **121**, 12233–12246 (2004).

[50] Schmitz, M. und P. Tavan. Vibrational spectra from atomic fluctuations in dynamics simulations. II. Solvent-induced frequency fluctuations at femtosecond time resolution. *J. Chem. Phys.* **121**, 12247–12258 (2004).

[51] Schmitz, M. und P. Tavan. On the art of computing the IR spectra of molecules in condensed phase. In S. Tanaka und J. Lewis (Herausgeber), *Modern methods for theoretical physical chemistry of biopolymers*, Kapitel 8, Seiten 157–177. Elsevier, Amsterdam (2006).

[52] Krimm, S. und Y. Abe. Intermolecular interaction effects in the amide I vibrations of β polypeptides. *Proc. Nat. Acad. Sci. USA* **69**, 2788–2792 (1972).

[53] Torii, H. und M. Tasumi. Model calculations on the amide-I infrared bands of globular proteins. *J. Chem. Phys.* **96**, 3379–3387 (1992).

[54] Torii, H. und M. Tasumi. 3-dimensional doorway-state theory for analyses of absorption bands of many-oscillator systems. *J. Chem. Phys.* **97**, 86–91 (1992).

[55] Torii, H. und M. Tasumi. Application of the 3-dimensional doorway-state theory to analyses of the amide-I infrared bands of globular proteins. *J. Chem. Phys.* **97**, 92–98 (1992).

[56] Lee, H., S. S. Kim, J. H. Choi und M. Cho. Theoretical study of internal field effects on peptide amide I modes. *J. Phys. Chem. B* **109**, 5331–5340 (2005).

[57] Torii, H. Time-domain calculations of the infrared and polarized Raman spectra of tetraalanine in aqueous solution. *J. Phys. Chem. B* **111**, 5434–5444 (2007).

[58] Krimm, S. und J. Bandekar. Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv. Prot. Chem.* **38**, 181–364 (1986).

[59] Hohenberg, P. und W. Kohn. Inhomogeneous electron gas. *Phys. Rev. B* **136**, 864–870 (1964).

[60] Kohn, W. und L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).

[61] Frisch, M. J., G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb *et al.* *Gaussian 98, Revision A.5*. Gaussian, Inc., 1998, Pittsburgh.

[62] Parr, R. G. und W. Yang. *Density-functional theory of atoms and molecules*. Oxford University Press, New York (1989).

[63] Perdew, J. und W. Yue. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Phys. Rev. B* **33**, 8800–8802 (1986).

[64] Troullier, N. und J. L. Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B* **43**, 1993–2006 (1991).

[65] Nonella, M. und P. Tavan. An unscaled quantum mechanical harmonic force field for p-benzoquinone. *Chem. Phys.* **199**, 19–32 (1995).

[66] Zhou, X., S. J. Mole und R. Liu. Density functional theory of vibrational spectra. 4. Comparison of experimental and calculated frequency of *all-trans*-1,3,5,7-octatetraene — The end of normal coordinate analysis? *Vib. Spectros.* **12**, 73–79 (1996).

[67] Neugebauer, J. und B. A. Hess. Fundamental vibrational frequencies of small polyatomic molecules from density-functional calculations and vibrational perturbation theory. *J. Chem. Phys.* **118**, 7215–7225 (2003).

[68] van Gunsteren, W. F. und H. J. C. Berendsen. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl.* **29**, 992–1023 (1990).

[69] Allen, M. P. und D. J. Tildesley. *Computer simulation of liquids*. Oxford University Press (1987).

[70] Verlet, L. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159**, 98–103 (1967).

[71] Ryckaert, J. P., G. Ciccotti und H. J. C. Berendsen. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkenes. *J. Comput. Phys.* **23**, 327–341 (1977).

[72] van Gunsteren, W. F. und H. J. C. Berendsen. Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.* **34**, 1311–1327 (1977).

[73] Ponder, J. W. und D. A. Case. Force fields for protein simulations. *Adv. Prot. Chem.* **66**, 27–85 (2003).

[74] Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan und M. Karplus. CHARMM: a Program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1983).

[75] Berendsen, H. J. C., D. van der Spoel und R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).

[76] Mathias, G., B. Egwolf, M. Nonella und P. Tavan. A fast multipole method combined with a reaction field for long-range electrostatics in molecular dynamics simulations: The effects of truncation on the properties of water. *J. Chem. Phys.* **118**, 10847–10860 (2003).

[77] Mathias, G. und P. Tavan. Angular resolution and range of dipole-dipole correlations in water. *J. Chem. Phys.* **120**, 4393–4403 (2004).

[78] Mathias, G. *Elektrostatische Wechselwirkungen in komplexen Flüssigkeiten und ihre Beschreibung in Molekulardynamiksimulationen*. Dissertation, Ludwig-Maximilians-Universität München (2004).

[79] Urey, H. C. und C. A. Bradley. The vibrations of pentatonic tetrahedral molecules. *Phys. Rev.* **38**, 1969–1978 (1931).

[80] MacKerell, Jr., A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck *et al*. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

[81] Stern, H., G. Kaminski, J. Banks, R. Zhou, B. Berne und R. Friesner. Fluctuating charge, polarizable dipole, and combined models: Parameterization from ab initio quantum chemistry. *J. Phys. Chem. B* **103**, 4730–4737 (1999).

[82] Rick, S., S. Stuart und B. Berne. Dynamical fluctuating charge force fields: applications to water. *J. Chem. Phys.* **101**, 6141–6156 (1994).

[83] Mitchell, P. und D. Fincham. Shell model simulations by adiabatic dynamics. *J. Phys.: Condens. Matter* **101**, 1031–1038 (1993).

[84] Vesely, F. N-particle dynamics of polarizable Stockmeyer-type molecules. *J. Comput. Phys.* **24**, 361–371 (1977).

[85] Applequist, J., J. Carl und K.-K. Fung. An atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *J. Am. Chem. Soc.* **94**, 2952–2960 (1972).

[86] Patel, S. und C. L. Brooks III. CHARMM fluctuating charge force field for proteins: I - Parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* **25**, 1–15 (2004).

[87] Patel, S., A. D. Mackerell und C. L. Brooks III. CHARMM fluctuating charge force field for proteins: II - Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.* **25**, 1504–1514 (2004).

[88] Lin, H. und D. Truhlar. QM/MM: what have we learned, where are we, and where do we go from there? *Theor. Chem. Acc.* **117**, 185–199 (2007).

[89] Eichinger, M., P. Tavan, J. Hutter und M. Parrinello. A hybrid method for solutes in complex solvents: Density functional theory combined with empirical force fields. *J. Chem. Phys.* **110**, 10452–10467 (1999).

[90] Grubmüller, H. und P. Tavan. Molecular dynamics of conformational substates for a simplified protein model. *J. Chem. Phys.* **101**, 5047–5057 (1994).

[91] Becker, O. M. und M. Karplus. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **106**, 1495–1517 (1997).

[92] Huisinga, W., C. Best, R. Roitzsch, C. Schütte und F. Cordes. From simulation data to conformational ensembles: structure and dynamics-based methods. *J. Comp. Chem.* **20**, 1760–1774 (1999).

[93] Giuliani, A. und C. Manetti. Hidden peculiarities in the potential energy time series of a tripeptide highlighted by a recurrence plot analysis: A molecular dynamics simulation. *Phys. Rev. E* **53**, 6336–6340 (1996).

[94] Daura, X., K. Gademann, B. Jaun, D. Seebach, W. van Gunsteren und A. Mark. Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed.* **38**, 236–240 (1999).

[95] Deuflhard, P., W. Huisinga, A. Fischer und C. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.* **315**, 39–59 (2000).

[96] Hamprecht, F. A., C. Peter, X. Daura, W. Thiel und W. F. van Gunsteren. A strategy for analysis of (molecular) equilibrium simulations: configuration space density estimation, clustering, and visualization. *J. Chem. Phys.* **114**, 2079–2089 (2001).

[97] Schütte, C., W. Huisinga und P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. In B. Fiedler (Herausgeber), *Ergodic theory, analysis, and efficient simulation of dynamical systems*, Seiten 191–224. Springer-Verlag (2001).

[98] Carstens, H. *Konformationsdynamik lichtschaltbarer Peptide: Molekulardynamiksimulationen und datengetriebene Modellbildung*. Dissertation, Ludwig-Maximilians-Universität München (2004).

[99] Altis, A., P. H. Nguyen, R. Hegger und G. Stock. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **126**, 244111:1–10 (2007).

[100] Chodera, J. D., N. Singhal, V. S. Pande, K. A. Dill und W. C. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **126**, 155101:1–17 (2007).

[101] Hinrichs, N. S. und V. S. Pande. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J. Chem. Phys.* **126**, 244101:1–11 (2007).

[102] Noe, F., I. Horenko, C. Schütte und J. C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.* **126**, 155102:1–17 (2007).

[103] Eckmann, J., S. Kamphorst und D. Ruelle. Recurrence plots of dynamical systems. *Europhys. Lett.* **4**, 973–977 (1987).

[104] Duda, R. O. und P. E. Hart. *Pattern Classification And Scene Analysis*. John Wiley and Sons (1973).

[105] Carstens, H., C. Renner, A. Milbradt, L. Moroder und P. Tavan. Multiple loop conformations of peptides predicted by molecular dynamics simulations are compatible with NMR. *Biochemistry* **44**, 4829–4840 (2005).

[106] Zadeh, L. A. *Fuzzy logic and its applications*. Academic Press (1965).

[107] Grauel, A. *Fuzzy-Logik*. BI Wissenschaftsverlag (1995).

[108] Dellnitz, M., A. Hohmann, O. Junge und M. Rumpf. Exploring invariant sets and invariant measures. *CHAOS: An Interdisciplinary Journal of Nonlinear Science* **7**, 221–228 (1997).

[109] Bishop, C. *Neural networks for pattern recognition*. Clarendon Press (1997).

[110] Schnell, P. Eine Methode zur Auffindung von Gruppen. *Biometr. Zeitschrift* **6**, 47–48 (1964).

[111] Parzen, E. On estimation of a probability density function and mode. *Annals Math. Statist.* **33**, 1065–1076 (1962).

[112] Hirschberger, T. *Hierarchische Klassifikation durch selbstorganisierende neuronale Algorithmen*. Diplomarbeit, Ludwig-Maximilians-Universität München, Lehrstuhl für Biomolekulare Optik, AG Theoretische Biophysik (2002).

[113] Rojas, R. *Neuronal networks: a systematic introduction*. Springer-Verlag, Berlin (1996).

[114] Ritter, H., T. Martinetz und K. Schulten. *Neuronale Netze*. Addison-Wesley (1990).

[115] Hillermeier, C., N. Kunstmann, B. Rabus und P. Tavan. Topological feature maps with self-organized lateral connections: a population-coded, one-layer model of associative memory. *Biol. Cybernetics* **72**, 103–117 (1994).

[116] Rabus, B. T. *Hypothesenbildung in phasenkodierten Assoziativspeichern*. Diplomarbeit, Technische Universität München, Physik-Department, Institut 30 (1992).

[117] Kohonen, T. *Self-Organization and Associative Memory*. Springer (1984).

[118] Dersch, D. R. und P. Tavan. Asymptotic level density in topological feature maps. *IEEE Trans. Neural Networks* **6**, 230–236 (1995).

[119] Dersch, D. R. *Eigenschaften neuronaler Vektorquantisierer und ihre Anwendung in der Sprachverarbeitung*. Dissertation, Ludwig-Maximilians-Universität München (1995).

[120] Gardiner, C. W. *Handbook of stochastic methods*. Springer-Verlag (1990).

[121] Reichl, L. E. *A modern course in statistical physics*. University of Texas Press (1980).

[122] Haken, H. *Synergetik: Eine Einführung*. Springer (1990).

[123] Kloppenburg, M. und P. Tavan. Deterministic annealing for density estimation by multivariate normal mixtures. *Phys. Rev. E* **55**, 2089–2092 (1997).

[124] Schultheis, V., R. Reichold, B. Schropp und P. Tavan. A polarizable force field for computing the infra-red spectra of the polypeptide backbone (2008). Zur Veröffentlichung in *J. Phys. Chem. B* angenommen.

[125] Yongye, A. B., J. Gonzalez-Outeiriño, J. Glushka, V. Schultheis und R. J. Woods. The conformational properties of methyl α-(2,8)-di/trisialosides and their N-acyl analogs: Implications for anti-Neisseria meningitidis B vaccine design (2008). Zur Veröffentlichung in *Biochemistry* eingereicht.

[126] Boatz, J. A. und M. S. Gordon. Decomposition of normal-coordinate vibrational frequencies. *J. Phys. Chem.* **93**, 1819–1826 (1989).

Danksagung

An dieser Stelle möchte ich allen danken, die zum Gelingen meiner Doktorarbeit beigetragen haben:

Allen voran Prof. Dr. Paul Tavan für die Idee zu diesem Projekt und für seine Betreuung.

Dem Boehringer Ingelheim Fonds für Seminare und für die Finanzierung, sowie allen seinen Mitarbeitern für ihr Interesse und für die persönliche Atmosphäre.

Den Sonderforschungsbereichen 533 und 749 der deutschen Forschungsgemeinschaft für finanzielle Unterstützung.

Thomas Hirschberger, der mir mit seinen Fragen oft geholfen hat.

Heiko Carstens, der mir die Serin-Tripeptid-Trajektorie zur Verfügung gestellt hat.

Rudolf Reichold für die Turbomolrechnungen von NMA.

Bernhard Schropp für die DFT/MM-Trajektorie von NMA in Wasser.

Gerald Mathias für seine Einführung in EGO.

Benjamin Rieff dafür, dass er mir angeboten hat, diese Arbeit Korrektur zu lesen, bevor ich das erste Wort geschrieben hatte. Außerdem Martin Lingenheil, Rudolf Reichold und Tobias Schrader für das Korrekturlesen.

Galina Babizki, Sebastian Bauer, Robert Denschlag, Bernhard Egwolf, Christine Lutz, Matthias Schmitz, Martina Stork, den bereits genannten Mitgliedern der Arbeitsgruppe und allen Diplomanden der Arbeitsgruppe für das kollegiale Klima.

Karl-Heinz Mantel und allen Systemadministratoren für ihre Mühen.

Frau Podolski, Frau Michaelis, Frau Widmann-Diermeier und Frau Haame für ihre Hilfe bei organisatorischen Fragen.

Den 'Experimentalos' für unzählige kulinarische Spaziergänge und dafür, dass sie mir ihre Sicht der (Infrarot-)Welt erklärt haben.

Austin Yongye und Robert Woods für die Zusammenarbeit.