

2.7 Berechnung des "Potential of Mean Force" durch "Umbrella Sampling"- Simulationen

"Potential of Mean Force"-Rechnungen beschäftigen sich unter anderem mit der Berechnung der Freien Energiedifferenz zwischen zwei Zuständen [Kirkwood 1935]. Die Grundlagen dieser Methode, wie in den folgenden Kapiteln vorgestellt, beschreiben immer Freie Energiekurven. Da in dieser Arbeit nur Untersuchungen an kondensierten Systemen durchgeführt wurden, kann immer näherungsweise angenommen werden, dass die beschriebenen Gleichungen auch für die Freien Enthalpiekurven gelten.

$$\Delta F^0 \approx \Delta G^0 \quad 2.28$$

In kondensierten Systemen kann angenommen werden, dass Änderungen der Freien Energie und Freien Enthalpie nahezu gleich sind.

2.7.1 Theoretische Grundlagen: "Potential of Mean Force"

Ein zentraler Punkt der statistischen Mechanik ist der Zusammenhang zwischen einer Potentialfunktion V (zum Beispiel dem *amber*-Kraftfeld), die die potentielle Energie eines Systems als Funktion ihrer Koordinaten (x_1 bis x_{3N}) beschreibt, und der Wahrscheinlichkeit P , das System in diesem Zustand anzutreffen:

$$P(x_1, x_2, x_3, \dots, x_{3N}) \propto e^{-V(x_1, x_2, x_3, \dots, x_{3N})} \quad 2.29$$

Bei "Potential of Mean Force"-Berechnungen wird die umgekehrte Vorgehensweise angewandt. Es wird ein Ausschnitt der Potentialfunktion aus einer beobachteten oder simulierten Wahrscheinlichkeitsverteilung bestimmt. Über den Zusammenhang zwischen Freier Energie und Zustandssumme eines Systems erhält man, wenn man einen einzelnen Freiheitsgrad aus der Integration extrahiert, eine Bestimmungsgleichung für das Freie Energieprofil entlang einer ausgewählten Reaktionskoordinate:

$$F^0 = -k T \ln Z(x_1, x_2, \dots, x_{3N}) \quad 2.30$$

$$Z = \int e^{\frac{-V(x_1, x_2, \dots, x_{3N})}{k T}} dx_1 dx_2 \dots dx_{3N} \quad 2.31$$

$$F^0(x_k) = -k T \ln \int e^{\frac{-V_{x_k}(x_1, x_2, \dots, x_{k-1}, x_{k+1}, \dots, x_{3N})}{k T}} dx_1 dx_2 \dots dx_{k-1} dx_{k+1} \dots dx_{3N} \quad 2.32$$

In Gleichung 2.32 ist x_k keine Variable von V mehr, sondern nur noch ein fester Parameter⁶⁵. Integriert man über alle verbleibenden Variablen, so erhält man eine Funktion, in der die Freie Energie des Systems an einem beliebigen Punkt der Reaktionskoordinate mit der Wahrscheinlichkeit, das System an diesem Punkt zu finden, zusammenhängt:

$$F^0(x_k) = k T (\ln P(x_k) + \ln C) \quad 2.33$$

Das Freie Energieprofil entlang x_k ließe sich also bis auf eine unbestimmte Integrationskonstante C bestimmen.

Um die Höhe und Form einer Energiebarriere zwischen zwei Zuständen A und B (Abbildung 26) zu bestimmen, könnte man mit Hilfe einer Moleküldynamik-Simulation einige Übergänge zwischen A und B simulieren und aus den Häufigkeiten, mit denen ein bestimmter Punkt auf der Reaktionskoordinate auftritt, dessen

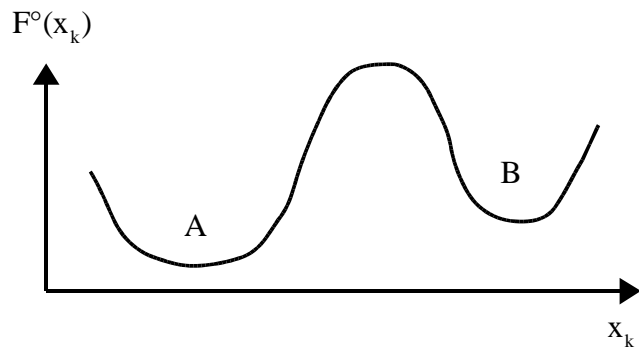


Abbildung 26 Energiebarriere zwischen zwei Zuständen A und B, die über die Reaktionskoordinate x_k ineinander übergehen. Über alle restlichen $3N-1$ Freiheitsgrade des Systems wurde per Integration summiert.

Lage im Freien Energieprofil bestimmen. Eine hinreichend lange

⁶⁵ Es ist aber immer noch eine Variable, von der die Freie Energie abhängt.

Simulationsdauer würde dabei sicherstellen, dass an jedem Punkt der Reaktionskoordinate alle relevanten Konformationen der restlichen Freiheitsgrade berücksichtigt werden. Dies entspricht der Integration über alle Koordinaten außer dem betrachteten Freiheitsgrad in Gleichung 2.33.

Das Problem dieser Vorgehensweise ist, dass bereits bei niedrigen Energiebarrieren nur wenige Übergänge zwischen A und B stattfinden. Das System befindet sich also nur selten im Bereich des Potentialmaximums zwischen A und B, so dass für eine statistisch signifikante Aussage zu wenige Werte in diesem Bereich vorliegen⁶⁶ (zu diesem Sachverhalt siehe auch "sampling problem", Seite 113). Dieses Problem lässt sich umgehen, indem man die Reaktionskoordinate in Teilbereiche aufspaltet und für jeden Teilbereich ein zusätzliches virtuelles Potential aktiviert. Da diese zusätzlichen Potentiale meist harmonisch (und damit parabelförmig) sind, spricht man auch von "umbrella sampling"⁶⁷.

"umbrella sampling"- Methode zur Abtastung von Reaktionskoordinaten

Die physikalische Potentialfunktion kann durch Addition eines zusätzlichen Potentials so modifiziert werden, dass im Laufe einer Simulation zuvor nur sehr schlecht abgetastete Bereiche der Reaktionskoordinate mit höherer Wahrscheinlichkeit erreicht werden [Kumar et al. 1992, Roux 1995] (Abbildung 27).

⁶⁶ Von dem Problem, dass selbst niedrige Energiebarrieren im Verlauf einer normalen Moleküldynamik- Simulation eines biochemischen Systems von einigen ns Dauer kaum überwunden werden können, und sich damit bei der Aufteilung des Systems auf die Zustände A und B nur extrem langsam ein Gleichgewicht einstellt, sei hier einmal abgesehen.

⁶⁷ Auch wenn die Ähnlichkeit zwischen Parabelkurve und Regenschirm ("umbrella") zweifelhaft ist, ist der Name der Methode fest etabliert, eine deutsche Übersetzung existiert nicht.

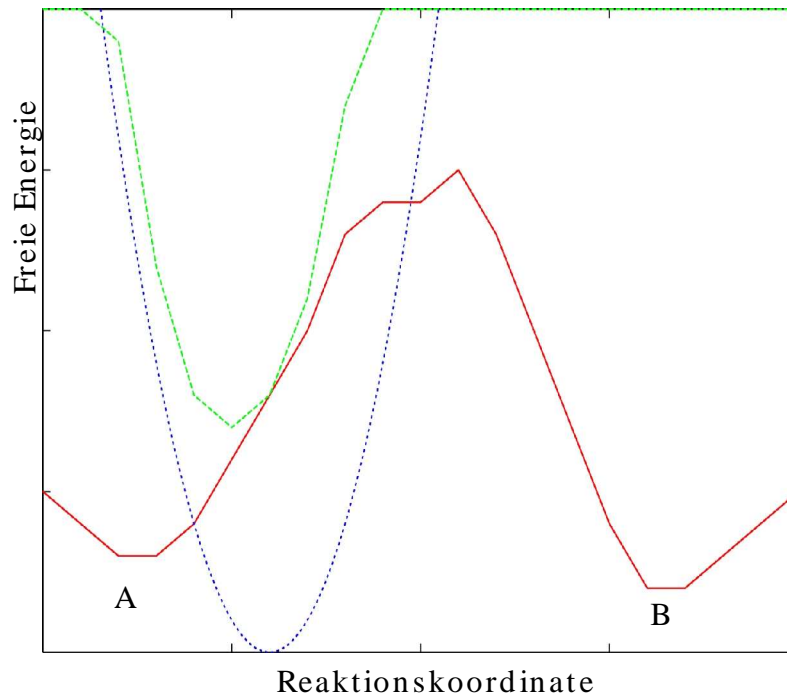


Abbildung 27 Effekt eines zusätzlichen harmonischen "umbrella"-Potentials auf die Form der Freien Energiekurve in willkürlichen Einheiten. Das physikalische Energieprofil mit zwei Minima (rot) wird durch Addition eines zusätzlichen harmonischen Potentials (blau) in ein Potential mit einem einzelnen verschobenen Minimum (grün) umgewandelt. Eine Simulation mit dieser modifizierten Potentialfunktion würde den Konformationsraum am Minimum der neuen Funktion effizient abtasten.

Ein hypothetisches ideales Zusatzpotential hätte die gleiche Form wie das physikalische Potential, nur horizontal gespiegelt. Damit würde man ein lineares Energieprofil erhalten, mit dem alle Punkte auf der Reaktionskoordinate mit gleicher Wahrscheinlichkeit erreicht werden. Da das physikalische Energieprofil aber nicht bekannt ist⁶⁸, verwendet man mehrere harmonische Potentiale und variiert die Lage von deren Minima (Abbildung 28):

⁶⁸ Ansonsten wäre die Durchführung einer "Potential of Mean Force"-Rechnung, die ja die Form des Energieprofils bestimmen soll, unnötig

$$V_i = V_0 + k \cdot (x - x_i)^2$$

2.34

Die modifizierte Potentialfunktion der i -ten Simulation. Zu der ursprünglichen Potentialfunktion V_0 wird ein harmonisches Potential einstellbarer Stärke k addiert. Das Minimum des zusätzlichen Potentials x_i beeinflusst den abgetasteten Bereich der Potentialfunktion. Im Verlauf der einzelnen Simulationsfenster würde x_i gleichmäßig über die Reaktionskoordinate verschoben.

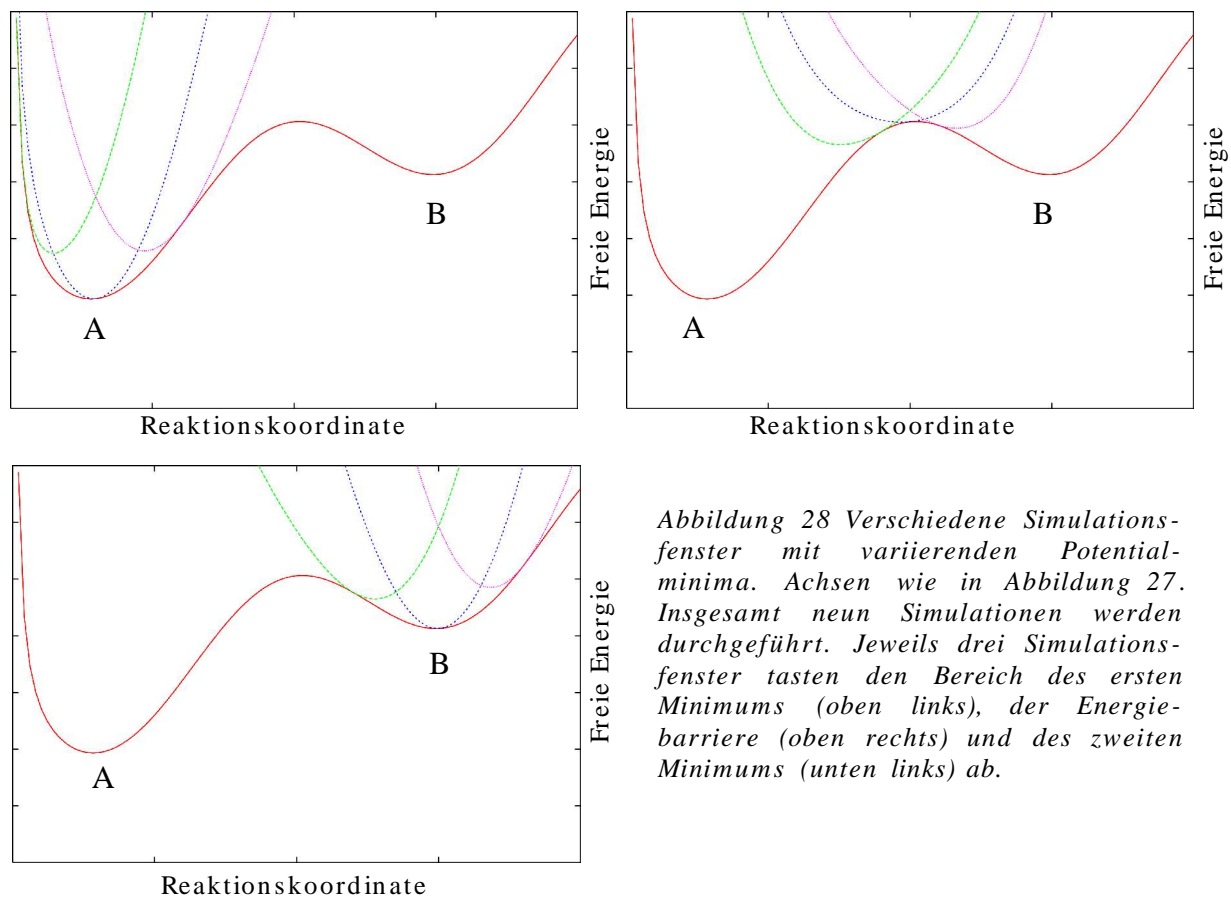


Abbildung 28 Verschiedene Simulationsfenster mit variierenden Potentialminima. Achsen wie in Abbildung 27. Insgesamt neun Simulationen werden durchgeführt. Jeweils drei Simulationsfenster tasten den Bereich des ersten Minimums (oben links), der Energiebarriere (oben rechts) und des zweiten Minimums (unten links) ab.

Jede einzelne dieser Simulationen tastet einen Bereich der Reaktionskoordinate nahe des Minimums des zusätzlichen Potentials ab.

Aus jedem Simulationsfenster lässt sich dann aus den Häufigkeiten, mit denen das System eine bestimmte Position auf der Reaktionskoordinate einnimmt, der Teil des Energieprofils nach Gleichung 2.33 bestimmen, den diese Simulation abgetastet hat. Dabei muss für die Berechnung der Wahrscheinlichkeit das zusätzliche, bekannte Potential mit berücksichtigt

werden. Da sich aus jedem Simulationsfenster nur ein Teilstück des Energieprofils inklusive einer unbestimmten Integrationskonstante bestimmen lässt, muss dafür gesorgt werden, dass sich die von den einzelnen Simulationsfenstern abgetasteten Bereiche ausreichend überlappen, um das Gesamtenergieprofil aus den Teilstücken eindeutig zusammensetzen zu können.

Üblicherweise würde man die Simulationen nacheinander durchführen, also als Startgeometrie der i -ten Simulation die Endgeometrie der $(i-1)$ -ten Simulation verwenden. Dadurch wird vermieden, dass Simulationen in unrealistischen Hochenergiekonformationen beginnen, was z.B. der Fall wäre, wenn man mit einem zusätzlichen Potential, das das Minimum der Gesamtpotentialfunktion in die Nähe des Zustands B verschiebt, die Simulation mit einer energieminierten Struktur im Zustand A beginnen würde. Um aus den in den verschiedenen Simulationsfenstern erhaltenen Häufigkeiten, das System an einem bestimmten Punkt der Reaktionskoordinate zu finden, ein Freies Energieprofil zu berechnen, sind verschiedene statistische Methoden möglich. In dieser Arbeit wurden die Häufigkeiten mit Hilfe des Programmes *wham* von Dr. A. Grossfield⁶⁹ nach der im Folgenden kurz vorgestellten WHAM-Methode⁷⁰ ausgewertet.

2.7.2 WHAM - Analyse von Häufigkeitsverteilungen aus "umbrella sampling"-Rechnungen

Um das Freie Energieprofil entlang einer Reaktionskoordinate nach Gleichung 2.32 zu bestimmen, muss die Wahrscheinlichkeitsverteilung $P(x)$ des Systems, wie in Gleichung 2.29, bekannt sein. Aus den Simulationsfenstern einer "umbrella sampling"-Simulation erhält man aber nur eine modifizierte Wahrscheinlichkeitsverteilung $P'(x)$, in die das zusätzliche harmonische Potential $V_i(x)$ mit einfließt. Das Freie Energieprofil des i -ten Simulationsfensters $F_i^o(x)$ ⁷¹ kann unter Berücksichtigung dieses

⁶⁹ *wham* ist erhältlich im WWW unter <http://dasher.wustl.edu/alan/wham/>.

⁷⁰ WHAM steht für **W**eighted **H**istogramm **A**nalysis **M**ethod, also Gewichtete Histogramm Analysemethode.

⁷¹ Das Freie Energieprofil hängt zwar im Idealfall nicht vom betrachteten Simulationsfenster ab, der Index soll aber verdeutlichen, dass jedes Simulationsfenster

Zusatzpotentials berechnet werden:

$$F_i^0(x) + V_i(x) = -k T \ln P_i'(x) + K_i \quad 2.35$$

Da die zusätzliche Integrationskonstante K_i nicht einzeln bestimmt werden kann, muss man unter Berücksichtigung aller Simulationsfenster einen Satz von K_i so bestimmen, dass die überlappenden Enden der Energieprofilteilstücke möglichst gut übereinstimmen. Da eine Simulation mit i Simulationsfenstern $i-1$ Überlappungsbereiche der Energieprofilteilstücke besitzt, bleibt dabei eine Konstante F unbestimmt übrig, so dass auch das gesamte Energieprofil nur relative Werte beinhaltet⁷².

Bei einer WHAM-Analyse werden die einzelnen K_i gleichzeitig mit der Wahrscheinlichkeitsverteilung ohne zusätzliche Potentiale $P(x)$ bestimmt. Die dazu nötigen Gleichungen wurden über eine ausführliche statistische Fehlerbetrachtung so aufgestellt, dass eine möglichst gute Näherung für $P(x)$ erhalten wird. Sie seien hier ohne Herleitung aus [Kumar et al. 1995] übernommen⁷³:

$$P(x) = \frac{\sum_i n_i(x)}{\sum_i N_i e^{-\frac{V_i(x) - K_i}{kT}}} \quad 2.36$$

$$K_i = -k T \ln \sum_x P(x) e^{-\frac{V_i(x)}{kT}} \quad 2.37$$

Die Auswertungsgleichungen für eine WHAM-Analyse. Der Laufindex i gibt an, dass über alle Simulationsfenster summiert wird. $P(x)$ ist die Wahrscheinlichkeitsverteilung ohne zusätzliches Potential, K_i sind die Integrationskonstanten der einzelnen Simulationsfenster. n_i ist die Häufigkeit, mit der das System an der Stelle x der Reaktionskoordinate aufzufinden ist. N ist die Zahl der Werte aus einer Simulation, $V_i(x)$ sind die zusätzlichen Potentiale der einzelnen Simulationsfenster. x bezeichnet einen bestimmten Bereich der Breite k auf der Reaktionskoordinate von x bis $x+k$, aufgrund von dieser Diskretisierung sind auch Summen- statt Integrationszeichen zulässig.

nur ein Teilstück F_i^0 des gesamtem Freien Energieprofils $F^0(x)$ bestimmt.

72 Ein absoluter Freier Energiewert wäre aufgrund eines fehlenden Nullpunktes der Freien Energieskala auch prinzipiell keine sinnvolle Angabe.

73 Eine Anpassung an die verwendete Nomenklatur, die sich an die Darstellung von Grossfield A. (Online- Publikation: http://dasher.wustl.edu/alan/talks/wham_talk.pdf) anlehnt, wurde vorgenommen.

Die Gleichungen 2.36 und 2.37 werden dann iterativ gelöst. Da aus einer Moleküldynamik-Simulation immer nur eine endliche Zahl diskreter Positionen des Systems entlang der Reaktionskoordinate erhalten werden können, kann auch nur eine diskrete Repräsentation der Wahrscheinlichkeitsverteilung bestimmt werden. Dazu teilt man den gesamten Bereich der Reaktionskoordinate in Teilstücke ein⁷⁴ und berechnet die mittlere Wahrscheinlichkeit, das System auf diesem Teilstück anzutreffen. Je mehr Daten aus der "umbrella sampling"-Simulation gewonnen wurden, desto kleiner kann die Breite dieser Teilstücke gewählt werden, und umso genauer wird damit die berechnete Wahrscheinlichkeitsverteilung $P(x)$. Aus $P(x)$ kann dann über Gleichung 2.33 das Freie Energieprofil und somit näherungsweise das Freie Enthalpieprofil berechnet werden.

⁷⁴ Deren Zahl und Breite muss nicht mit der der Simulationsfenster übereinstimmen, sie sollten aber nicht wesentlich kleiner sein, um statistische Fehler zu vermeiden.