Theoretische Molekulare Biophysik

Zusammenfassung/Übersicht

erstellt von Tizian Römer

basierend auf der Vorlesung
Theoretische Molekulare Biophysik
von Prof. Dr. Wolfgang Welzel
aus dem Wintersemester 2018/2019
am Karlsruher Institut für Technologie (KIT)

Kontakt via E-Mail: tiroemer@yahoo.de

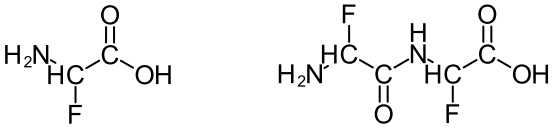
Inhaltsverzeichnis

1 Proteine und Aminosäuren

1.1 Aminosäuren

PEPTIDBINDUNG:

Proteine und Peptide (kleine Proteine) sind Polymere aus Aminosäuren. 21 Aminosäuren kommen natürlich in ähnlicher Häufigkeit vor. Sie alle sind von der Form



wobei F für einen Rest steht, der sich für die 21 Aminosäuren unterscheidet. Unter Abspaltung von H_2O kann sich die Aminogruppe NH_2 mit der Hydroxylgruppe COOH einer anderen Aminosäure verbinden (Peptidbindung). So kann sich eine Kette von Aminosäuren bilden, die dann ab etwa 30 Aminosäuren als Protein bezeichnet wird (sonst als *Peptid*). Proteine sind typischerweise 30 bis 400 Aminosäuren lang.

EIGENSCHAFTEN VON AMINOSÄUREN:

Die 21 Aminosäuren lassen sich aufteilen in

- 3 positiv geladene,
- 2 negativ geladene,
- 4 polare,
- 8 hydrophobe und
- 4 sonstige Aminosäuren.

Es gibt keine kovalenten Bindungen zwischen den Resten (außer den Schwefel-Bindungen des Cysteins); Proteine sind daher lineare Polymere.

PROTEIN DATA BANK:

In der Protein-Datenbank stehen etwa 100 000 3D-Proteinstrukturen und 10 000 000 Gensequenzen zur Verfügung.

ZAHLENWERTE:

Räumliche Größen

– eines Atoms:	1 Å
– einer C — C-Bindung:	1,4 Å

Energien

– Raumtemperatur (kT):	0,6 kcal
– C — C-Bindung:	140 kcal/mol
– Elektrostatische WW:*	5 kcal/mol
– Dipol-Dipol-WW:	0,3 kcal/mol
– Wasserstoffbrückenbind.:	3 — 4 kcal/mol
– Hydrophobische WW mit H2O:	2 — 3 kcal/mol
– Schwefelbrücken:	80 kcal/mol

(*bei Wasser-Dielektrizität $\epsilon \approx 80$)

Bindungen bis etwa 10 kcal/mol sind unter Raum-/Körpertemperaturbedingungen nicht vollständig stabil, sondern können sich bilden und lösen. Diese Bindungen stabilisieren die Proteinstruktur.

STATISTISCHE MECHANIK EINER BINDUNG:

Die Wahrscheinlichkeit, dass eine Bindung geschlossen ist, beträgt

$$p_{\text{geb}} = \frac{e^{-\beta E_{\text{geb}}}}{Z} = \frac{e^{-\beta E_{\text{geb}}}}{e^{-\beta E_{\text{geb}}} + e^{-\beta E_{\text{ung}}}} = \frac{1}{1 + e^{-\beta(E_{\text{ung}} - E_{\text{geb}})}}.$$

1.2 Experimentelle Bestimmung der Proteinstruktur

KRISTALLOGRAPHIE:

Gelingt es, viele Proteine einer Sorte in einem regelmäßigen Gitter anzuordnen, kann an dieser Anordnung per Röntgenstrahlung Bragg-Reflexionen beobachtet werden. Die Anzahl an Bragg-Peaks ist proportional zur Anzahl an Atomen in einer Elementarzelle, also zur Anzahl an Atomen in einem Protein.

Je größer ein solcher Kristall ist (d. h. je mehr Proteine er enthält), desto größer ist die Intensität des gestreuten Lichts. Die Herstellung von Protein-Kristallen ausreichender Größe ist schwierig. Man versucht daher, die einfallende Intensität zu erhöhen; wodurch der Kristall jedoch im Femto-/Nanosekundenbereich explodiert. Um kein durch die Explosion verschmiertes Bild zu bekommen, darf nur ein sehr kurzer, intensiver Lichtpuls auf den Kristall treffen.

KERNSPINRESONANZ (NMR):

Bei der Kernspinresonanz werden die Proteine mit elektromagnetischen Wellen bestrahlt. Die Kerne ^{1}H, ^{13}C und ^{15}N haben durch ihren Spin ein kleines magnetisches Moment, das in Resonanz mit der eingestrahlten Welle liegen kann. Die genaue Absorptionsenergie ΔE wird dabei geringfügig von den Nachbaratomkernen beeinflusst. Aus dieser Verschiebung im Absorptionsspektrum lässt sich dann auf die Molekülstruktur rückschließen.

In einer zweidimensionalen Kernspinresonanz werden gleichzeitig zwei Frequenzen eingestrahlt. Die diagonale des entsprechenden Flächendiagrams entspricht der eindimensionalen Kernspinresonanz. Abseits der diagonale gibt es ebenfalls Peaks, die sich jeweils zwei Diagonal-Peaks zuordnen lassen. Diese Verbindung entsteht durch Kopplungen der entsprechenden Atome.

CRYO ELECTRON MICROSCOPY:

Auch mit Elektronenmikroskopen lassen sich Proteinstrukturen auflösen. Vorteile sind eine hohe Auflösung (bis 3 Å) und dass auch die innere Proteinstruktur zugänglich ist. Das nötige Vakuum erschwert jedoch die Präparation der Probe, die zudem besonders dünn sein muss, damit nicht alle Elektronen absorbiert werden. Und auch hier werden die Proben von den Elektronenstrahlen zerstört, sodass nur kurze Messungen möglich sind.

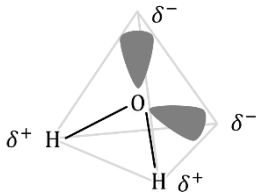
ABBÉ-LIMIT:

Die Auflösung eines Objekts in einem Medium mit Brechungsindex n durch Licht der Wellenlänge λ und einem halben Öffnungswinkel des Objektivs α beträgt

$$d = \frac{\lambda}{n \sin \alpha}.$$

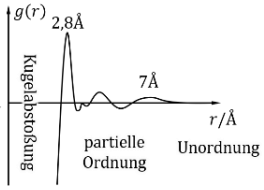
1.3 Löslichkeit

WASSERSTOFFBRÜCKENBINDUNGEN:



Das Wasserstoffatom trägt die Partialladungen $\delta^{\pm} = \pm 0{,}325\, e$. Die vier Partialladungen sind in den Ecken eines Tetraeders angeordnet.

Durch diese regelmäßige Anordnung können sich Strukturen wie das Eis-Gitter herausbilden.

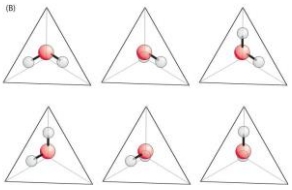


Die Abstandskorrelationsfunktion $g(r)$ für die Verteilung von Wassermolekülen hat einen Peak bei 2,8 Å. In diesem Abstandsbereich befinden sich im Schnitt

$$N = \int_{2\text{Å}}^{3\text{Å}} 4\pi r^2\, dr\; g(r) \approx 3 \text{ bis } 4$$

andere Wassermoleküle.

ENTROPIE UND HYDROPHOBER EFFEKT:



Es gibt insgesamt

$$\binom{4}{2} = 6$$

verschiedene Ausrichtungen für ein Wassermolekül in einem Tetraeder (4 über 2: vier Ecken und eine Spiegelsymmetrie). Man betrachte Wasser an einer hydrophoben Wand. Angenommen, eine Tetraederecke ist dadurch hydrophob. Dadurch bleiben nur noch

$$\binom{3}{2} = 3$$

verschiedene Ausrichtungen. Der Unterschied in der Entropie ist demnach

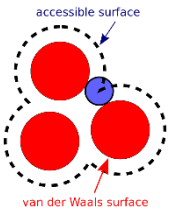
$$T\,\Delta S = kT \ln \frac{3}{6} \approx -0{,}7kT.$$

Für $kT \approx 0{,}6$ kcal/mol (siehe 1.1) und zehn Wassermolekülen pro nm^2 folgt für die Entropie pro Fläche

$$\frac{T\,\Delta S}{A} = -0{,}6 \text{ kcal} \cdot 0{,}7 \cdot 10 \text{ nm}^{-2} \approx -4 \text{ kcal nm}^{-2}.$$

Im Experiment erhält man $\Delta S_{\text{exp}}/A \approx -2$ kcal nm^{-2}. Selbst unsere sehr einfache Abschätzung liefert also von der Größenordnung her sinnvolle Ergebnisse. Die durch die hydrophobe Wand verringerte Entropie sorgt wegen $F = U - TS$ für eine erhöhte freie Energie; daher ist die Lösung hydrophober Moleküle energetisch unvorteilhaft bzw. aggregieren sich hydrophobe Partikel, um die Oberfläche zu Minimieren. Dies bezeichnet man als *Hydrophoben Effekt*.

SOLVENT ACCESSIBLE SURFACE AREA (SASA):



Die von einem Lösungsmittel erreichbare Oberfläche SASA eines Moleküls wird genähert durch:

$$\text{SASA} = \sum_i 4\pi R_i^2 - \sum_{\langle ij \rangle} \text{S}_{ij} + \sum_{\langle ijk \rangle} S_{ijk} + \cdots,$$

wobei R_i der Abrollradius (gestrichelte Linie) *der einzelnen Atome* des Moleküls ist. S_{ij} ist die Überlappungsfläche zweier Benachbarter Atome und S_{ijk} die Überlappung der Überlappungsflächen dreier benachbarter Atome usw. (höhere Ordnungen werden in der Regel vernachlässigt).

Experimentell kann man zeigen, dass die Lösungsentropie tatsächlich proportional zur SASA ist: Das heißt, man kann folgende Moleküle bzgl. Löslichkeit folgendermaßen anordnen:

$$\text{CH}_4 > \text{C}_2\text{H}_6 > \text{C}_3\text{H}_8 > \cdots.$$

In einer Anordnung bzgl. der Entropiedifferenz ΔS müsste man die $> \rightarrow <$ entsprechend umdrehen.

BORN-MODELL:

Man betrachte eine homogene, kugelförmige Verteilung der Ladung q mit Radius R. Das von ihr erzeuge elektrische Feld hat die Energie (mit $\epsilon = \epsilon_0 \epsilon_r$)

$$U = \frac{1}{2}\int_R^\infty d^3r\, \epsilon E^2 = \frac{1}{2}\int_R^\infty 4\pi r^2\, dr\, \epsilon \left(\frac{q}{4\pi\epsilon r^2}\right)^2 = \frac{q^2}{8\pi\epsilon R}.$$

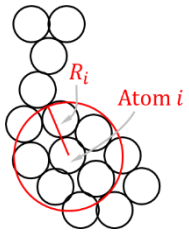
Die Energie, die nötig ist, um die Verteilung von einer Umgebung mit Dielektrizität ϵ_1 nach ϵ_2 zu bringen ist daher

$$\Delta U = \frac{q^2}{8\pi R}\left(\frac{1}{\epsilon_2} - \frac{1}{\epsilon_1}\right) = -\frac{q^2}{8\pi R}\left(1 - \frac{1}{\epsilon}\right) > 0,$$

wobei zuletzt $\epsilon_2 = \epsilon$ sowie $\epsilon_1 = 1$ gesetzt wurde, unter der Annahme, dass die Ladungsverteilung von Vakuum/Luft in ein Medium ϵ überführt wird.

Die Idee dieser Herleitung von ΔU ist, dass die Energie, um ein Ion in ein Lösungsmittel zu geben, dieselbe ist, die nötig ist, um ein Ion zu entladen, ungeladen in ein Lösungsmittel zu überführen und dort wieder aufzuladen.

VERALLGEMEINERTES BORN-MODELL:



Für ein nicht homogenes und/oder kugelförmiges Molekül lässt sich Born-Modell leicht verallgemeinern:

$$\Delta U_{\text{Molekül}} = \sum_{\text{Atome } i} \Delta U(R_i).$$

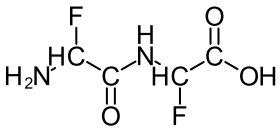
Dabei werden einfach die Lösungsenergien des Born-Modells aller Atome aufsummiert. R_i ist dabei der *effektive Born Radius*, also im Prinzip der Abstand des Atoms zum Lösungsmittel.

2 Sekundärstruktur und Faltung

2.1 Ramachandran-Plot

DIEDER-/DIHEDRALWINKEL:

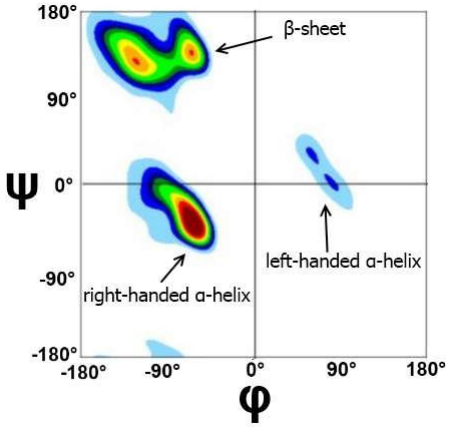
Die Peptidbindung ist Planar, das heißt alle Atome zwischen zwei Kohlenstoffatomen, an denen ein Rest F hängt, liegen in einer Ebene.



Zwei solche Ebenen, die jeweils an einem C-Atom mit Rest F hängen, können jedoch gegeneinander verdreht sein. Legt man eine Ausrichtung durch die Richtung der Bindung zum Rest F fest, ergeben sich relativ dazu die Winkel *Dieder-* oder *Dihedralwinkel* ϕ, ψ für die beiden anhängenden Peptidbindungsebenen.

DER RAMACHANDRAN-PLOT:

Experimentelle Daten darüber, wie häufig bestimmte Winkel ϕ, ψ in der Natur auftauchen, kann man in einem sogenannten *Ramachandran-Plot* darstellen.



Es gibt einige Winkelkonstellationen, die sehr viel häufiger vorkommen als andere. In letzteren sind sich womöglich die Reste F im Weg. Erstere tauchen in häufig anzutreffenden immergleichen Strukturen auf, wie α-Helices und β-Faltblätter.

2.2 α-Helices und β-Faltblätter

In jeder Peptidbindung gibt es eine CO- und eine NH-Gruppe. Das O ist stark elektronegativ, das H elektropositiv. In einer Aminosäurenkette gibt es zwei Konformationen derart, dass sämtliche CO- und NH-Gruppen der Kette jeweils einander gegenüberliegen, um sich gegenseitig zu stabilisieren. Diese Strukturen sind die α-Helices und β-Faltblätter.

Da jede Aminosäure die nötigen CO- und NH-Gruppen hat, können alle Aminosäuren an α-Helices und β-Faltblättern mitwirken (Ausnahme: Prolin; hier ist die NH-Gruppe ausnahmsweise in einen Ring eingebaut). Durchschnittlich 70% aller Aminosäuren in einem Protein sind Teil einer α-Helix oder eines β-Faltblatts.

2.3 Primär-, Sekundär-, Tertiär-, Quartärstruktur

Primärstruktur:

Die reine Aminosäuresequenz.

Sekundärstruktur:

Lokale Strukturen mit einigen wenigen Aminosäuren: α-Helices und β-Faltblätter.

Tertiärstruktur:

Komplette Konformation eines Proteins; Anordnung der verschiedenen Sekundärstrukturen relativ zueinander.

Quartärstruktur:

Komplexe aus mehreren Proteinen.

2.4 Anfinsen-Experiment und Levinthal-Paradoxon

ANFINSEN-EXPERIMENT:

Urea ist ein Molekül um nicht-kovalente Verbindungen zu zerstören, β-mercaptoethanol zerstört Schwefel-Brücken. Unter Zugabe beider Stoffe werden native (funktionsfähig konformierte) Proteine denaturiert. Entzieht man die Stoffe wieder, falten sich die Proteine wieder in den nativen Zustand. Entzieht man *erst* β-mercaptoethanol und *anschließend* Urea , schließen sich falsche Schwefelbrücken, und der native Zustand wird nicht wieder erreicht. Nicht-kovalente Wechselwirkungen tragen also wesentliche zur Konformation bei. Gibt man nun (nach Entzug auch von Urea) wieder ein wenig β-mercaptothenaol hinzu, brechen die falsch geschlossenen Schwefelbrücken auf und können sich nun unter Abwesenheit von Urea auch wieder korrekt schließen.

LEVINTHAL-PARADOXON:

Selbst wenn jede Aminosäure nur zwei Faltungszustände annehmen könnte und eine Änderung der Konformation 10^{-13} s dauern wurde, bräuchte ein Protein mit 150 Aminosäuren

$$2^{150} \cdot 10^{-13}\ \text{s} > 10^{24}\ \text{Jahre},$$

um aus allen Möglichkeiten die native Konformation zu finden.

LÖSUNG DES PARADOXON – ZERKLÜFTUNG DER ENERGIE:

Jedes Protein hat abhängig von seiner Primärstruktur eine bestimmte Energielandschaft $E(\text{Zustand})$ im Zustandsraum. Für eine zufällige Primärstruktur ist diese stark zerklüftet und es ist unmöglich, das globale Minimum zu finden – hier schlägt das Levinthal-Paradoxon zu. Es gibt aber bestimmte Primärstrukturen, deren Energielandschaft nicht so zerklüftet, sondern eher trichterförmig ist: Hier ist es sehr viel wahrscheinlicher, im globalen Grundzustand zu landen. Primärstrukturen natürlicher Proteine haben diese Eigenschaft – bei der Evolution der Proteine war also nicht nur die Funktion des Proteins in einer bestimmten Konformation wichtig, sondern auch ganz einfach die Faltbarkeit.

Ein Maß für die Zerklüftung R einer Energielandschaft kann definiert werden als

$$R(Q) = \sum_{i,j} \delta_{Qq_{ij}} p_i p_j, \qquad p_i = e^{-\beta E_i}/Z,$$

wobei i und j Zustände/Konformationen bezeichnet und p_i die Wahrscheinlichkeit dafür ist, dass sich das System im i-ten Zustand befindet. q_{ij} ist der Abstand der Zustände im (diskreten) Zustandsraum.

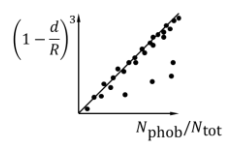
2.5 Micelle-Modell der Proteinstruktur

Das Micelle-Modell der Proteinstruktur geht davon aus, dass sich außen am Protein eine Schicht von hydrophilen Aminosäuren befindet, während sich die hydrophoben Aminosäuren im Inneren befinden (siehe 1.3). Für ein kugelförmiges Protein mit Radius R und einer Schicht hydrophilen Aminosäuren der Dicke d als Kugelschale beträgt das Volumen des hydrophoben Anteils

$$V_{\text{phob}} = \frac{4}{3}\pi(R-d)^3$$

bzw. dessen Anteil am Gesamtvolumen

$$\frac{V_{\text{phob}}}{V_{\text{tot}}} = \frac{\frac{4}{3}\pi(R-d)^3}{\frac{4}{3}\pi R^3} = \left(1-\frac{d}{R}\right)^3 = \frac{N_{\text{phob}}}{N_{\text{tot}}},$$



wobei das Volumenverhältnis dem Anzahlverhältnis entspricht, wenn alle Aminosäuren etwa gleich groß sind. Trägt man die linke Seite auf die y-Achse und die rechte auf die x-Achse auf, erfüllen alle Punkte auf der Einheitsgerade diese Gleichung. Die meisten realen Proteine liegen auf dieser Gerade, was das Micelle-Modell bestätigt. Proteine, die nicht auf der Gerade liegen, sind solche, die sich mit anderen Proteinen zu Quartärstrukturen zusammensetzen und daher einen höheren hydrophoben Anteil haben.

3 Molekulare Mechanik

3.1 Klassisches Kraftfeld/Potential

GRUNDLAGEN:
Atomare Interaktionen sind von quantenmechanischen Effekten dominiert; deren Berechnung ist jedoch sehr aufwändig. Zum Verständnis von Funktion und Vorhersage der Konformation von größeren Molekülen verwendet man daher ein klassisches Kraftfeld $\vec{F}(\vec{x}) - \nabla V(\vec{x})$, wobei $\vec{x}$ ein Vektor aus allen $3N$ kartesischen Koordinaten der N Atome darstellt. Es gibt $3N - 6$ Freiheitsgrade (die 6 Freiheitsgrade der Translation/Rotation Rotation des gesamten Moleküls sind uninteressant). Bindungen (1–2) und Winkel (1–3) sind sehr starr, die Schwankungen machen typischerweise nur 1 % bzw. 2 % aus.

BINDUNG (1–2):
Zwei Atome, die kovalent aneinandergebunden sind, tragen zum Potential den folgenden Term bei

$$V_B = k_B(x - x_0)^2$$

bei, mit Gleichgewichtsabstand x_0 und Steifigkeit k_B.

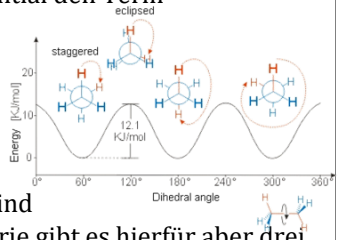
WINKEL (1–3):
Ein Winkel zwischen zwei benachbarten kovalenten Bindungen trägt zum Potential den Term

$$V_A = k_A(\theta - \theta_0)^2$$

bei. Ein solcher Winkel wird durch zwei benachbarte Bindungen, also drei benachbarte Atome definiert.

DIHEDRAL-WINKEL (1–4):
Ein Dihedral-Winkel ist ein Winkel zwischen zwei Flächen, die jeweils von drei Atomen definiert werden. Zwei benachbarte Flächen können sich zwei Atome teilen, sodass pro Fläche ein weiteres Atom die Ausrichtung definiert; insgesamt braucht es zur Definition eines Dihedral-Winkels also vier Atome. Ein solcher Winkel trage dann zum Potential den Term

$$V_D = \sum_n k_{D,n}\big(1 - \cos\big(n(\phi - \phi_0)\big)\big).$$



n ist die *Multiplizität*: Zum Beispiel ist bei Ethan die energieärmste Konformation, wenn die beiden H_3-Enden gegeneinander verdreht sind *(„staggered“)*. Aufgrund der Symmetrie gibt es hierfür aber drei Möglichkeiten. Die Summe geht hier also über $n = 1, 2, 3$.

VAN-DER-WAALS:
Je zwei Atome i und j üben van-der-Waals-Wechselwirkungen aufeinander aus, die in der Form

$$V_W = k_W\left(\left(\sigma_{ij}/r_{ij}\right)^{12} - 2\left(\sigma_{ij}/r_{ij}\right)^{6}\right).$$

Die Proportionalität $\sim -1/r^6$ lässt sich quantenmechanisch nachrechnen. Die 12 ist willkürlich gewählt, sodass sie sich durch Quadrierung aus dem r^{-6}-Term ergibt (numerisch schnell zu rechnen) und sorgt für einen steilen Anstieg für kleine r_{ij}. Der Faktor 2 sorgt dafür, dass das Minimum dieses Potentials genau bei σ_{ij} liegt. Das Potential fällt schnell ab, sodass dieser Term nur für nah benachbarte Atome berücksichtigt werden muss.

ELEKTROSTATIK:
Das elektrostatische Potential zwischen zwei Atomen ist

$$V_{ES} = \frac{q_i q_j}{4\pi\epsilon r_{ij}}.$$

Dabei wird $\epsilon_r = \epsilon/\epsilon_0$ normalerweise als $3 - 5$ gewählt; Wasser wäre $\epsilon_r = 80$, hydrophobe Umgebungen haben etwa $\epsilon_r = 3$. Da die Verbindungslinien zwischen zwei Atomen vor allem über den (oft hydrophoben) Molekülbereich verlaufen, liegt ϵ_r näher am hydrophoben Wert als am Wasser. Das elektrostatische Potential fällt nur langsam mit r_{ij}^{-1} ab; es ist daher der einzige Term, dessen Berechnungsdauer quadratisch in der Anzahl der Atome geht. Er verursacht in der Praxis die größten Probleme.

3.2 Modelle und Programmierung der Dynamik

MODELLE FÜR PARAMETER:
Es gibt verschiedene Modelle wie AMBER, CHARM oder GROMOS, die sich darin unterscheiden, wie die Parameter k_X, x_0, θ_0 etc. gewählt werden. Sie entstanden aus quantenmechanischen Berechnungen oder aus Messungen, wurden jedoch unsystematisch iterativ angepasst, um bestmögliche Übereinstimmungen mit dem Experiment zu erzielen. Je nach chemischer Umgebung kann ein bestimmtes Atom unterschiedliche Parameter haben; so kennt AMBER zum Beispiel 13 verschiedene Parameter für ein Kohlenstoffatom. Entsprechend gibt es noch viel mehr verschiedene Bindungstypen, je nach Art der Bindungspartner.

ERWEITERUNGEN:
Der nächstgrößte Effekt, der standardmäßig nicht mehr berücksichtigt wird, sind durch die Umgebung induzierte Partialladungen, ähnlich wie bei den van-der-Waals-Kräften. In diesem Fall könnte man die Atomladungen abhängig von ihrer Umgebung anpassen:

$$q_i \to q_i + \delta q_i(\text{Umgebung}).$$

DYNAMIK:
Aus dem Potential mit all diesen Termen ergibt sich ein Kraftfeld, sodass sich dann die Newtonschen Bewegungsgleichungen numerisch lösen lassen. Die Verlet-Integrationsvorschrift lautet

$$\vec{r}(t + \Delta t) = \vec{r}(t) + \vec{v}(t)\,\Delta t + \frac{1}{2m}\vec{F}(t)\,\Delta t^2,$$
$$\vec{v}(t + \Delta t) = \vec{v}(t) + \frac{1}{2m}\Big(\vec{F}(t) + \vec{F}(t + \Delta t)\Big)\Delta t.$$

Man berechnet also zunächst $\vec{r}(t + \Delta t)$, woraus sich $\vec{F}(t + \Delta t)$ ergibt. Anschließend erst berechnet man $\vec{v}(t + \Delta t)$, da man so die mittlere Kraft aus den Zeiten t und $t + \Delta t$ nutzen kann und sich die Geschwindigkeit dadurch schneller updated.

DIE ZEITSCHRITTE:
Ein Zeitschritt sollte dabei deutlich kleiner sein, als die kleinste Periodendauer einer Molekülschwingung, die sich aus den k_B bzw. k_A ergibt. Sagen wir, ein Zeitschritt soll ein Zehntel davon betragen,

$$\Delta t = \frac{T_{\min}}{10}.$$

Eine C − C-Bindung hat eine Eigenfrequenz von etwa 10^{14} Hz. Also braucht man $\Delta t \approx 10^{-15}$ s. Beste Simulationen erreichen 10^{12} Schritte und simulieren entsprechend rund 10^{-3} s.

THERMO- UND BARIOSTAT:
Die Impulse $\vec{p}_i$ werden nach jedem Zeitschritt so reskaliert, dass Druck oder Temperatur

$$P = \frac{1}{V}\sum_{j>i}\left(F_{ij}\left(|\vec{q}_i - \vec{q}_j|\right) + \frac{|\vec{p}_i|}{m_i}\right), \quad kT = \frac{1}{3N}\sum_i \frac{|\vec{p}_i|^2}{m_i} = \frac{2}{3}\langle E_{\text{kin}}\rangle$$

konstant bleiben.

COARSE GRAINING:
Da auch mit Parallelprogrammierung kürzere Laufzeiten schwierig sind, versucht man eher, Strukturen (mehrere Atome) zusammenzufassen und mit effektiven Wechselwirkungen zu rechnen.

3.3 Strukturbasierte Modelle

Strukturbasierte Modelle können zum Einsatz kommen, wenn die native Konformation eines Proteins bereits bekannt ist; man nutzt die Information über die bekannte native Konformation zur besseren Simulation aus. Man kann diese Modelle dann nutzen, um etwas über den Faltungs*prozess* herauszufinden oder über andere Eigenschaften (z. B. Umschalten zwischen zwei Konformationszuständen).

KONTAKTE UND CONTACT ORDER:

Die intramolekularen Kontakte zwischen entfernten Aminosäuren, die das Protein konformiert halten, sind dann bekannt. Man definiert die *Contact Order* eines Proteins als

$$\mathrm{CO} = \frac{1}{LN}\sum_{i<j} \Delta s_{ij},$$

wobei L die Proteinlänge (also die Anzahl Aminosäuren), N die Kontakt-Anzahl und Δs_{ij} der Sequenz-Abstand zweier Kontakte i, j ist. Eine höhere Contact Order geht in der Regel mit einer längeren Faltungszeit einher. Eine niedrige Contact Order ist ein Charakteristikum von Downhill-Folding-Proteinen (siehe 3.4).

MD-POTENTIAL MIT KONTAKT-ANZIEHUNGSTERM:

Der Einfachheit halber nehmen wir an, dass nur native Kontakt-Wechselwirkungen möglich/erlaubt sind, also nur solche, die auch im nativen Zustand vorkommen.

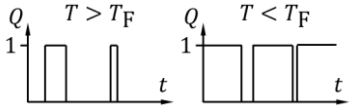
Wir können nun das Physik-basierte Potential aus 3.1 manipulieren; auf atomarer Ebene berücksichtigen wir nun nur die Bindungs-, Winkel- und Dihedral-Winkel-Terme aus 3.1. Statt die van-der-Waals- und elektrostatischen Kräfte auf atomarer Ebene zu betrachten, fügen wir nur ein Potential für die bekannten nativen Kontakte hinzu,

$$V = \cdots + \sum_{\text{Kontakte } i,j} k_K \left(\left(\kappa_{ij}/r_{ij}\right)^{12} - 2\left(\kappa_{ij}/r_{ij}\right)^{6} \right),$$

sowie einen Term der Hard-Core-Abstoßung, damit nach wie vor keine Atome aufeinander liegen:

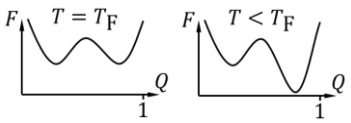
$$V = \cdots + \sum_{\text{Atome } i,j} k_{HC}\left(\sigma_{HC}/r_{ij}\right)^{12}.$$

Für den Anteil der geschlossenen Kontakte Q gilt dann, dass in der Regel $Q = 0$ odr $Q = 1$ ist – das Protein ist meisten entweder komplett ent- oder gefaltet. Die Faltungstemperatur T_F ist als diejenige Temperatur definiert, bei der $\langle Q \rangle = 1/2$ gilt.



3.4 Konkurrenz zwischen Entropie und Innerer Energie

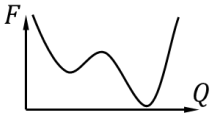
Gefaltete Zustände haben eine geringe Innere Energie, dafür haben entfaltete eine geringe Entropie (d. h. es gibt sehr viel mehr unterschiedliche entfaltete als gefaltete Zustände). Die Funktion der Inneren Energie $F(Q)$ mit Anteil geschlossener Kontakte Q (siehe 3.3) sieht daher typischerweise wie folgt aus:



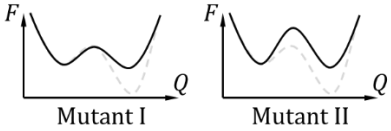
$Q = 1$ ist wegen der sehr hohen Entropie wieder relativ unwahrscheinlich; die Funktionsfähigkeit des Proteins ist auch schon ab einem relativ hohen Faltungsgrad $Q > Q_\mathrm{cr}$ mit $Q_\mathrm{cr} < 1$ gegeben. Im eher seltenen Fall von *Downhill-Folding*-Proteinen gibt es kein lokales Maximum zwischen den Minima.

3.5 Phi-Wert

Man betrachte ein Protein mit folgender $F(Q)$-Kurve bei $T < T_\mathrm{F}$:



Eine Mutation (Wechsel einer für die Faltung relevanten Aminosäure) kann diese Kurve nun folgendermaßen verändern:



Man stelle sich vor, dass das Protein beim Übergang vom entfalteten in den gefalteten Zustand eines Zwischenzustand mit $Q \approx 1/2$ einnimmt, der auf dem lokalen Maximum liegt. Wenn die Mutation in einem Bereich des Proteins stattfindet, der außerhalb des im Zwischenzustand gefalteten Bereichs liegt, wird $F(1/2)$ nicht verändert (Mutation I). Liegt die Mutation innerhalb des im Zwischenzustand bereits gefalteten Bereichs, wird auch $F(1/2)$ verändert (typischerweise verschlechtert/ erhöht).

Man definiert für eine Mutation den ϕ-Wert

$$\phi := \frac{F(\mathrm{ZS}) - F'(\mathrm{ZS})}{F(\mathrm{N}) - F'(\mathrm{N})} \in [0, 1],$$

wobei ZS für Zwischenzustand, also $Q \approx 1/2$, und N für nativen Zustand, also $Q \approx 1$, steht. Die gestrichenen sind die neuen freien Energien nach der Mutation. Es gilt in der Regel

$$F(\mathrm{N}) - F'(\mathrm{N}) < 0$$

und daher

$$\text{Mutation I:} \quad F(\mathrm{ZS}) - F'(\mathrm{ZS}) \approx 0 \implies \phi \approx 0$$
$$\text{Mutation II:} \quad F(\mathrm{ZS}) - F'(\mathrm{ZS}) < 0 \implies \phi \approx 1.$$

4 Monte Carlo

4.1 Grundlagen

STOCHASTISCHER PROZESS UND ERWARTUNGSWERT:

Man betrachte ein System aus einer Menge von Zuständen $\{x_i\}$, wobei i ein Index über alle möglichen Zustände ist. Die Idee ist, einen stochastischen Prozess zu generieren, der eine Sequenz von Konfigurationen des Systems $\{X_t\}$ liefert, wobei in jedem Zeitschritt t das System einen bestimmten Zustand x_i einnimmt. Wenn das System zum Beispiel bei $t = 35$ im Zustand Nr. $i = 354$ ist, gilt $X_{35} = x_{354}$.

Wenn die Wahrscheinlichkeit für den i-Zustand $p(x_i)$ sei, dann gilt für eine beliebige Observable θ der Erwartungswert

$$\langle\theta\rangle = \frac{1}{Z}\sum_i \theta(x_i)\, e^{-\beta E(x_i)} = \sum_i \theta(x_i)p(x_i).$$

MARKOV-PROZESS UND MASTERGLEICHUNG:

In einem Markov-Prozess hängt der Folgezustand X_{t+1} nur vom aktuellen Zustand X_t ab. Zwischen allen möglichen Zuständen gibt es Übergangswahrscheinlichkeiten W_{ij} von einem Zustand x_i in einen Zustand x_j.

Sei P_{it} die Wahrscheinlichkeit, dass sich das System nach t Zeitschriten im Zustand x_i befindet (das heißt, dass $X_t = x_i$ ist). Die Änderung der Wahrscheinlichkeit nach der Zeit t im i-ten Zustand zu sein, entspricht der Wahrscheinlichkeit $\sum_j W_{ji}P_{j,t-1}$ aus einem anderen Zustand in den i-ten Zustand zu welchseln abzüglich der Wahrscheinlichkeit $\sum_j W_{ij}P_{i,t-1}$ aus dem Zustand i in einen anderen Zustand zu wechseln:

$$P_{i,t+1} - P_{it} = \sum_j \left(W_{ji}P_{jt} - W_{ij}P_{it}\right).$$

Diese Gleichung heißt *Mastergleichung*.

DETAILED BALANCE:

Für $t \to \infty$ sollte sich das System möglichst in einem Gleichgewicht einpendeln, sodass sich P_{it} für große Zeiten nicht mehr verändert: $P_{i,t+1} = P_{it}$ für große t. Die obige Summe muss also verschwinden. Eine Möglichkeit hierzu ist *Detailed Balance*: Alle Summenglieder verschwinden individuell:

$$W_{ji}P_{jt} = W_{ij}P_{it}.$$

Im Gleichgewicht, für große t, kann man annehmen, dass P_{it} eine Boltzmann-Verteilung ist: $P_{i,t\to\infty} = e^{-\beta E(x_i)}$. Man sollte die W_{ij} also so wählen, dass für dieses $P_{i,t\to\infty}$ *Detailed Balance* erfüllt ist.

METROPOLIS-KRITERIUM:

Eine mögliche Wahl für W_{ij} ist wie folgt:

$$W_{ij} = C \cdot \begin{cases} e^{-\beta(E_j - E_i)}, & E_j > E_i, \\ 1, & E_j \le E_i, \end{cases} \quad E_i := E(x_i), \quad C = \text{const.}$$

Wenn die Energie des „neuen" Zustands also geringer ist, soll er auf jeden Fall angenommen werden. Ist die Energie des neuen Zustands größer, kann er dennoch angenommen werden, wobei die Wahrscheinlichkeit mit zunehmender Energiedifferenz abnimmt.

4.2 Helix-Stabilität

ISING-MODELL-ÄHNLICHE BETRACHTUNG:

Man betrachte ein Protein mit N Aminosäuren. Eine Aminosäure ist im Zustand 1, wenn sie in einer Helix verbaut ist und im Zustand 0, wenn sie nicht in einer Helix verbaut ist. Ein Zustand eines Proteins wird also dargestellt als Folge von Einsen und Nullen. Zwei benachbarte Aminosäuren können drei verschiedene Energien haben:

E_{00}, falls beide nicht in einer Helix verbaut sind,
E_{10}, falls eine verbaut ist und eine nicht und
E_{11}, falls beide verbaut sind.

Der Einfachheit halber kann man $E_{00} = 0$ setzen. Die interessante Observable ist der Helix-Anteil des Proteins

$$\theta := \frac{\langle n\rangle}{N} = \frac{1}{N}\sum_n n\, p(n) = \frac{1}{NZ}\sum_i n_i\, e^{-\beta E_i}$$

wobei $\langle n\rangle$ die mittlere Anzahl an Aminosäuren ist, die in einer Helix verbaut ist. $p(n)$ ist die Wahrscheinlichkeit dafür, dass n Aminosäuren in einer Helix verbaut sind. Die Summe über i sind alle möglichen Zustände, $Z = \sum_i e^{-\beta E_i}$ ist die Zustandssumme.

PROTEIN MIT EINER HELIX:

Wir betrachten nun nur Zustände mit *einer* Helix, also mit nur *einer* zusammenhängenden Kette von Einsern. Die Energie ist dann, falls $E_{00} = 0$,

$$E(n) = \tilde\theta(n)(2E_{10} + (n-1)E_{11}), \quad \tilde\theta(n) := \begin{cases} 0, & n = 0, \\ 1, & n > 0. \end{cases}$$

Dabei vernachlässigen wir, dass es nur ein Aminosäurenpaar mit Energie E_{01} gibt, falls die Helix am Rand der Kette liegt. Für die Zustandssumme folgt, da es in einer Kette von N Aminosäuren

$$g(n) := \begin{cases} 1, & n = 0, \\ N - n + 1, & n > 0. \end{cases}$$

Möglichkeiten gibt, eine zusammenhängende Kette der Länge n anzuordnen,

$$\begin{aligned} Z &= \sum_{n=0}^{N} g(n)\, e^{-\beta E(n)} = 1 + \sum_{n=1}^{N} g(n)\, \underbrace{e^{-2\beta E_{10}}}_{=:\sigma}\, \underbrace{e^{-\beta(n-1)E_{11}}}_{=:s^{n-1}} \\ &= 1 + \sigma\sum_{n=1}^{N} g(n)\, s^{n-1} = 1 + \sigma\sum_{n=0}^{N-1}(N-n)s^n \\ &= 1 + \sigma N\sum_{n=0}^{N-1} s^n - \sigma s\sum_{n=0}^{N-1} n s^{n-1}. \end{aligned}$$

Somit folgt

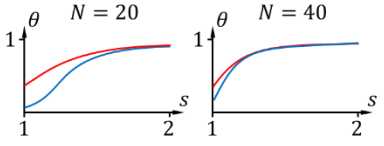
$$Z = 1 + \sigma N f(s) - \sigma s f'(s), \qquad f(s) := \frac{1 - s^N}{1 - s}$$

und ferner

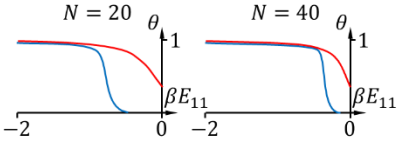
$$\begin{aligned} \theta &= \frac{1}{NZ}\sum_{n=0}^{\infty} g(n)\, n\, e^{-\beta E(n)} = \frac{1}{NZ}\frac{\partial}{\partial s}\sigma\sum_{n=1}^{\infty} g(n)\, s s^{n-1} \\ &= \frac{1}{NZ}\frac{\partial}{\partial s}s(Z-1) = \frac{1}{N}\left(1 - \frac{1}{Z} + \frac{s}{Z}\frac{\partial Z}{\partial s}\right) \approx \frac{1}{N}\frac{s}{Z}\frac{\partial Z}{\partial s} = \frac{1}{N}\frac{\partial \ln Z}{\partial \ln s}. \end{aligned}$$

Dabei wurde genähert, dass aus $Z > 1$ folgt, dass $1 - 1/Z < 1$ ist und bei der durchschnittlichen Anzahl von in Helix eingebauten Aminosäuren vernachlässigt werden kann.

PLOTS:



$\sigma = 0{,}1$ $\sigma = 0{,}001$

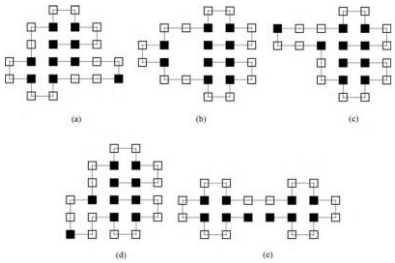


$\beta E_{01} = 1$ $\beta E_{01} = 8$

4.3 HP-Modell

ANNAHMEN:

Im HP-Modell wird nur zwischen hydrophoben (H) und polaren (P) Aminosäuren unterschieden. Alle Aminosäuren haben dieselbe Größe und hängen an einer Kette, die in einem Gitter angeordnet wird:



Die Energie einer Konformation hänge dann nur von den nächsten Nachbarn (im Gitter, nicht nur in der Kette) ab:

$$E = \sum_{\langle ij \rangle} \epsilon(x_i, x_j),$$

wobei ϵ davon abhängt, ob die Gitterpunkte x_i, x_j polar (P), hydrophob (H) oder leer (E) sind, etwa:

$$\epsilon = \begin{pmatrix} & \mathbf{P} & \mathbf{H} & \mathbf{L} \\ \mathbf{P} & 0 & >0 & <0 \\ \mathbf{H} & & 0 & >0 \\ \mathbf{L} & & & 0 \end{pmatrix}$$

(und symmetrisch für die untere linke Hälfte).

CODIERUNG DER KONFORMATION:

Die Konformation wird durch $N-1$ Buchstaben R = right, S = straight und L = left codiert (die erste Aminosäure bekommt keinen Buchstaben). Ein Zustand ist also zum Beispiel durch RRLLSSLR⋯ charakterisiert. Die Lage im Raum (Rotation) ist damit automatisch irrelevant. Somit gibt es $\sim 3^{N-1}$ Konformationen (tatäschlich etwas weniger, da Überlappungen verboten sind), es besteht also wieder das Levinthal-Paradoxon. Ein kleiner Änderungsschritt der Konformation bedeutet nicht nur den Austausch eines Buchstabens, da dadurch die Lage großer Teile der Kette verändert wird; man darf immer nur einen Punkt der Kette verschieben, wozu zwei Buchstaben verändert werden müssen.

STABILITÄT EINES GEFALTETEN ZUSTANDS:

Die Wahrscheinlichkeit im i-ten Zustand zu sein, beträgt

$$p_i = \frac{e^{-\beta E_i}}{Z} = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}} \cdot \frac{e^{\beta E_0}}{e^{\beta E_0}} = \frac{e^{-\beta(E_i - E_0)}}{1 + \sum_{j \neq 0} e^{-\beta(E_j - E_0)}}.$$

Wir betrachten einen Zustand i als gefaltet, falls $p_i \sim 1$ ist; wenn es einen solchen Zustand gibt, dann den Grundzustand $i = 0$:

$$p_0 = \frac{1}{1 + \sum_{j \neq 0} e^{-\beta(E_j - E_0)}} > \frac{1}{1 + 3^{N-1} e^{-\beta(E_1 - E_0)}}.$$

Damit $p_0 \sim 1$ gilt, brauchen wir also

$$3^{N-1} e^{-\beta(E_1 - E_0)} \sim 1 \qquad \Longrightarrow \qquad E_1 - E_0 \gg kT.$$

Zudem muss der Grundzustand eindeutig sein.

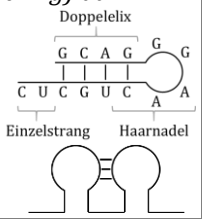
METROPOLIS MONTE CARLO SIMULATION:

Man beginne bei einem beliebigen Zustand (Konformation) und wähle einen zufälligen neuen Zustand. Unter einer bestimmten Bedingung (z. B. Metropolis-Kriterium, siehe 4.1), die nur vom aktuellen und neuen Zustand abhängt, wechsele man vom aktuellen Zustand in den neuen; ist die Bedingung nicht erfüllt, behalte man den aktuellen Zustand bei und wähle einen anderen zufälligen neuen Zustand usw.

5 RNA-Faltung

5.1 Grundlagen – Sekundär- und Tertiärstruktur

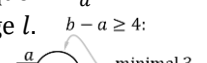
Die Konformation von RNA basiert größtenteils auf Basenpaaren $G-C$ und $A-U$; das heißt die einsträngige RNA faltet sich so, dass sie mit sich selbst passende Basenpaare ausbilden kann. Ein einzelnes Basenpaar ist instabil; zusätzliche Stabilität erhält die RNA durch Stacking: Zwei verbundene Teilstränge eines RNA-Einzelstrangs bilden wie DNA ein helixartige Struktur aus, die auch durch vertikale Wechselwirkungen *(Stacking)* dann stabilisiert wird. An Sekundärstruktur gibt es in erster Linie ebendiese Doppelhelix, sowie lose, einzelsträngige RNA. Typisch ist dann auch eine Haarnadelstruktur. Eine Tertiärstruktur bildet sich ebenfalls durch Basenpaare-Verbindungen zwischen Sekundärstrukturen aus.

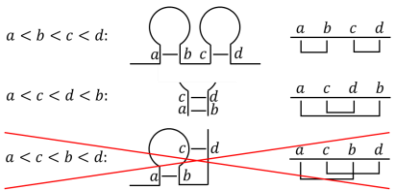


5.2 RNA-Strukturvorhersage – Formalisierung

Die meisten RNA-Studien beziehen sich nur auf die Sekundärstruktur. Input-Daten ist die Basensequenz $B_a \in \{G, C, A, U\}$ mit $a = 1, \dots, l$ bei Kettengesamtlänge l.

Ein Basenpaar ist nur erlaubt, falls die beiden beteiligten Basen B_a, B_b mindestens vier Basen auseinander liegen, das heißt $b - a \geq 4$. Ein Haarnadelende enthält also mindestens drei Basen. Für zwei Basenpaare $B_a - B_b$ und $B_c - B_d$ können die vier Indizes drei verschiedene Reihenfolgen haben. Die letzte Möglichkeit, sogenannte *Pseudoknots*, fällt unter die Tertiärstruktur; diese werden bei der Sekundärstrukturvorhersage nicht erlaubt.





5.3 RNA-Strukturvorhersage – Algorithmus

DYNAMISCHE PROGRAMMIERUNG:

Am einfachsten versieht man jedes gebildete Basenpaar mit einer Energie $-\varepsilon$ und Loopregionen betrachtet man nicht als Nachteil. Letztlich geht es also nur darum, die Konformation mit den meisten Basenpaaren zu finden.

ALGORITHMUS:

Sei B_{ab} die maximal mögliche Anzahl an gleichzeitig geschlossenen Basenpaaren innerhalb eines RNA-Abschnitts zwischen (inklusive) den Basen a und b.

Wie bestimmen wir B_{ab}? Zunächst ist $B_{ab} = 0$, falls $b - a < 4$ (siehe 5.2). Dies schließt auch ein, dass $B_{ab} = 0$, falls $a > b$.

Falls nun aber a und b weit genug auseinander liegen, unterscheiden wir die folgenden beiden Möglichkeiten:

Entweder geht die letzte Base b *keine* Bindung innerhalb des Abschnitts ein: Dann ist $E_{ab} = E_{a,b-1}$.

Oder die letzte Base b geht eine Bindung mit irgendeiner Base c ein, wobei $c \in [a, b-4]$ sein muss, um den nötigen Abstand von 4 Positionen zu gewährleisten (siehe 5.2). In diesem Fall erhält man also ein Basenpaar, allerdings sind weitere Paare in den Abschnitten a bis $c-1$ und $c+1$ bis $b-1$ möglich.

Diese Vorschriften können wir ganz allgemein wie folgt notieren:

$$B_{ab} = \begin{cases} 0, & b - a < 4, \\ \max \begin{cases} B_{a,b-1} \\ \max\limits_{\substack{a \leq c \leq b-4 \\ \text{und } P_{cb} \text{ erlaubt}}} \{1 + B_{a,c-1} + B_{c+1,b-1}\}, \end{cases} & \text{sonst,} \end{cases}$$

wobei P_{cb} dann erlaubt ist, wenn die Basen c und b gemäß der Regel $G-C$ und $A-U$ eine Paarung eingehen *können.* Falls P_{cb} für kein c erlaubt ist, ergibt sich natürlich eine für das Maximum über c.

5.4 RNA-Strukturvorhersage – Händische Berechnung

DAS BEISPIEL:

Wir betrachten als Beispiel die Primärstruktur

GGGAAAAUCCAU

und schreiben B_{ab} als Koeffizienten einer Matrix.

SOFORTIGE VEREINFACHUNG DES ALGORITHMUS:

Wir können die Matrix in vier Bereiche einteilen:

- Da $B_{ab} = 0$ für $b - a < 4$ ist der ganz dunkle Teil der Matrix automatisch immer null.
- Da $c \geq a$ gilt, folgt $(b-1) - (c+1) \leq b - a - 2$. Somit gilt $B_{c+1,b-1} = 0$ noch bis zur $4 + 2 = 6$-ten Diagonale.
- Da $c \leq b - 4$ gilt, folgt $(c-1) - a \leq b - a - 5$. Somit gilt $B_{a,c-1} = 0$ noch bis zur $4 + 5 = 9$-ten Diagonale.
- Erst ab der 10. Diagonale können alle Terme beitragen.

AUSFÜLLEN DER MATRIX:

Gemäß unserer Rekursionsvorschrift hängt B_{ab} stets nur von solchen B_{cd} mit $c \geq a$ und $d < b$ ab – also von Einträgen, die weiter unten links in der Matrix liegen. Daher füllen wir die Matrix Diagonale für Diagonale von unten links nach oben rechts

$$B_{15} \to B_{26} \to \cdots \to B_{6,10} \quad \to \quad B_{16} \to B_{27} \to \cdots.$$

	b	1	2	3	4	5	6	7	8	9	10	11	12	
a		G	G	G	A	A	A	A	U	C	C	A	U	
1	G	0	0	0	0	0	0	0	1	2	3	3	3	
2	G	0	0	0	0	0	0	0	1	2	3	3	3	
3	G	0	0	0	0	0	0	0	1	2	2	2	2	
4	A	0	0	0	0	0	0	0	1	1	1	1	1	
5	A	0	0	0	0	0	0	0	0	0	0	0	1	09. Diag.
6	A	0	0	0	0	0	0	0	0	0	0	0	1	
7	A	0	0	0	0	0	0	0	0	0	0	0	1	
8	U	0	0	0	0	0	0	0	0	0	0	0	0	06. Diag.
9	C	0	0	0	0	0	0	0	0	0	0	0	0	
10	C	0	0	0	0	0	0	0	0	0	0	0	0	
11	A	0	0	0	0	0	0	0	0	0	0	0	0	
12	U	0	0	0	0	0	0	0	0	0	0	0	0	

Da $B_{1,12} = 3$, sind maximal drei Basenpaare möglich.

BEISPIEL-ELEMENT:

Zum Beispiel berechnen wir das Element B_{29} wie folgt:

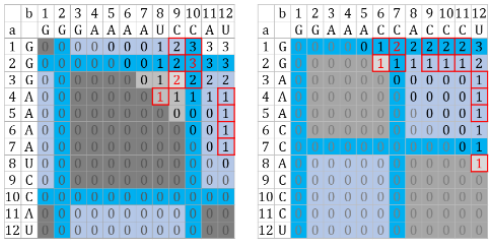
$$B_{29} = \max \begin{cases} B_{28} \\ \max\limits_{\substack{2 \leq c \leq 5 \\ \text{und } P_{c9} \text{ erlaubt}}} \{1 + B_{2,c-1} + B_{c+1,8}\} \end{cases}$$

$$= \max \begin{cases} 1 \\ 1 + B_{21} + B_{38} \\ 1 + B_{22} + B_{48} \end{cases} = \max \begin{cases} 1 \\ 1 + 0 + 1 = 2 \\ 1 + 0 + 1 \end{cases}$$

Da die Base $b = 9$ ein C ist, gilt $P_{c9} =$ erlaubt mit $2 \leq c \leq 5$ nur für $c = 2, 3$, da nur dort ein G ist ($c = 4, 5$ ist ein A).

SCHLUSS AUF DIE GESCHLOSSENEN BASENPAARE:

Wir kennen nun die maximale Anzahl an Basenpaaren, aber wissen noch nicht, *welche* Basenpaare wir schließen müssen, um genau diese Anzahl zu erhalten.

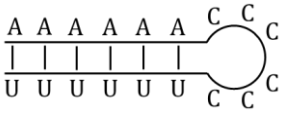


Dazu sind in den obigen beiden Beispielen die gemäß $C-G$ und $A-U$ möglichen Basenpaare rot umrandet. Aus den umrandeten Kästchen wähle man dasjenige (a, b) mit der größten Zahl aus, das der Matrixdiagonale am nächsten liegt (hier: $(2, 10)$ bzw. $(1, 7)$); seine Zahl erhält eine rote Farbe. Anschließend färbe man die Zeilen und Spalten a und b blau; man erhält ein Quadrat. Alles, was links/rechts und über/unter dem Quadrat liegt, wird violett gefärbt. Diese Paare sind aufgrund von 5.2 verboten. Mit dem Inhalt des Quadrats und den diagonal dazu liegenden, grauen Flächen, gehe man rekursiv vor, wie zu Beginn mit der gesamten Matrix. Die roten Zahlen sind dann die geschlossenen Basenpaare.

5.5 Kinetik der RNA-Faltung

EINFACHE SEQUENZ:

Man betrachte im einfachsten Fall einen Haarnadelpin der Sequenz $A_6C_6U_6$:



Im Wesentlichen sind bei dieser Sequenz entweder alle oder keine der A − U-Bindungen geschlossen.

KINETISCHE EIGENSCHAFTEN:

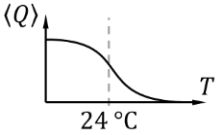
Diese beiden Konformationen bilden ein Gleichgewicht mit Übergangsraten k_g bzw. k_e:

$$\text{entfaltet} \underset{k_e}{\overset{k_g}{\rightleftarrows}} \text{gefaltet}.$$

Experimentell findet man, dass $k_g \approx 10^4\ s^{-1}$ ist und zwar unabhängig von der Temperatur, wohingegen

$$k_e \in [10^3, 10^5]\ s^{-1} \quad \text{für} \quad T \in [4, 34]\ °C.$$

Auch wenn eine einzelne RNA zu einem bestimmten Zeitpunkt entweder ganz gefaltet oder entfaltet ist, findet man für den (Ensemble- oder Zeit-)Mittelwert Q einen kontinuierlichen Temperaturverlauf:

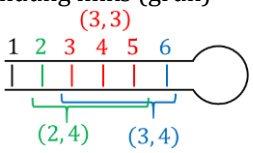


EINFACHES MODELL:

Nehmen wir an, wir können den Zustand der RNA $A_6C_6U_6$ mit den folgenden Parametern beschreiben:

$n \in [0, 6]$: Anzahl geformter Basenpaare,
$m \in [1, 7 - n]$: Position des ersten Basenpaares.

Wir können nun die Bindungszustände über Tupel (m, n) beschreiben. Befinden wir uns etwa im Zustand (m, n) (rot), dann können wir entweder eine weitere Bindung links (grün) oder rechts schließen (blau):



rechts: $(m, n) \to (m, n + 1)$
link: $(m, n) \to (m - 1, n + 1)$

In beiden Fällen sinkt die Freie Energie durch ein zusätzliches Basenpaar sowie durch Stacking; auf der anderen Seite gibt es auch einen (geringfügigen) Entropieverlust. Beim Bilden des ersten Basenpaares allerdings ist der Entropieverlust sehr viel größer und der Energiegewinn durch Stacking bleibt aus. Nehmen wir mal an, dass sich immer erst die Bindung 6 bildet, dann 5 usw. Dann bekommen wir folgendes Gleichgewichtsschema:

$$\text{entfaltet} \underset{k_{-l}}{\overset{k_l}{\rightleftarrows}} (6,1) \underset{k_-}{\overset{k_+}{\rightleftarrows}} (5,2) \underset{k_-}{\overset{k_+}{\rightleftarrows}} \cdots \underset{k_-}{\overset{k_+}{\rightleftarrows}} (1,6) = \text{gefaltet},$$

wobei $k_{\pm l}$ die Rate ist, mit der ein Loop und ein Basenpaar gebildet/geöffnet wird und $k_\pm$ die Rate, mit der ein Basenpaar hinzukommt/bricht. Da die Barriere für das erste Basenpaar deutlich größer als, als für die weiteren, können wir annehmen, dass die Faltungsrate in etwa der Loop-Rate entspricht:

$$k_g \approx k_l \sim e^{-\beta\, \Delta E_{\text{Loop}}} = 1/N_{\text{Loop}},$$

wobei wir den experimentellen Befund $\Delta E_{\text{Loop}} = kT \ln N_{\text{Loop}}$ eingesetzt haben. N_{Loop} ist die Loopgröße, also bei uns $N_{\text{Loop}} = 6$ (die 6 C-Basen). Somit ist k_g temperaturunabhängig. Zudem gilt

$$k_e = \frac{k_{-l}}{N_{\text{STEN}}} \underbrace{\left(\frac{k_-}{k_+}\right)^{N_{\text{STEN}}-1}}_{\sim \exp(-\beta\, \Delta E_{\text{STEN}})},$$

womit k_e für steigende Temperaturen zunimmt. N_{STEN} steht für die Länge der Helix (in unserem Fall also auch 6).

6 Free Energy Methods

6.1 Freie Energien

Freie Energien sind wichtige Kenngrößen thermodynamischer Systeme und gleichzeitig in Simulationen nicht einfach zu berechnen. Aus der Statistischen Mechanik ist bekannt, dass sich die Freie Energie durch

$$F = -kT \ln Z\,, \qquad Z = \int d\Gamma\, e^{-\beta E(\Gamma)}$$

berechnen lässt, wobei Γ der komplette (hochdimensionale) Phasen- bzw. Zustandsraum darstellt. In der Regel sind in Simulationen Differenzen Freier Energien (bei $T = \text{const}$)

$$\Delta F = -kT \ln \frac{\int d\Gamma\, e^{-\beta E_2(\Gamma)}}{\int d\Gamma\, e^{-\beta E_1(\Gamma)}}$$

relevanter und leichter zu berechnen.

6.2 Wahrscheinlichkeitsverteilung und Neugewichtung

WAHRSCHEINLICHKEITSVERTEILUNG UND HISTOGRAMME:

Sei $\mathcal{P}(E)$ die Wahrscheinlichkeitsverteilung von bspw. der Energie E. In einer Simulation kann man $\mathcal{P}$ durch ein Histogramm $h(E)$ mit Binbreite δE darstellen, in dessen Bins die n Simulationsergenisse je nach Energie E einsortiert werden. Sei $\mathcal{H}$ das normierte Histogramm, dann gilt

$$\mathcal{H}(E) \coloneqq \frac{h(E)}{n\,\delta E}, \qquad \lim_{\delta E \to 0, n \to \infty} \mathcal{H}(E) = \mathcal{P}(E).$$

Auf diese Weise können wir aus einer Simulation $\mathcal{P}(E)$ näherungsweise ableiten.

VERBINDUNG ZU STATISTISCHEN GRÖSSEN:

Mit einer Zustandsdichte $g(E)$ können wir

$$\mathcal{P}(\Gamma) = \frac{e^{-\beta E(\Gamma)}}{Z}, \qquad \mathcal{P}(E) = \frac{g(E) e^{-\beta E}}{Z} = \frac{e^{\mathcal{S}(E) - \beta E}}{Z},$$

$$Z = \int d\Gamma\, e^{-\beta E(\Gamma)} = \int dE\, g(E)\, e^{-\beta E} = \int dE\, e^{\mathcal{S}(E) - \beta E}$$

schreiben, wobei wir die dimensionslose Entropie

$$\mathcal{S}(E) \coloneqq \ln g(E)$$

eingeführt haben. Aus $\mathcal{S}$ folgt dann etwa die mittlere Energie

$$\langle E \rangle = \int dE\, E\, \mathcal{P}(E) = \frac{\int dE\, E\, e^{\mathcal{S}(E) - \beta E}}{\int dE\, e^{\mathcal{S}(E) - \beta E}}.$$

Und $\mathcal{S}$ können wir aus einem simulierten Histogramm $\mathcal{P}$ ableiten: Invertieren der obigen Formel liefert

$$\mathcal{S}(E) = \ln Z \mathcal{P} e^{\beta E} = \ln \mathcal{P} + \beta E - \beta F.$$

Man beachte, dass $\mathcal{S}$ nicht von T abhängt. Die T-Abhängigkeit in $\mathcal{P}$ und F müssen sich genau aufheben. Aus obiger Formel ergibt sich direkt eine Formel für die Differenz von Freien Energien bei verschiedenen Temperaturen (aber gleicher Energie E):

$$\beta_2 F(T_2) - \beta_1 F(T_1) = \ln \frac{\mathcal{P}(E, T_2)}{\mathcal{P}(E, T_1)} + (\beta_2 - \beta_1) E.$$

NEUGEWICHTUNG:

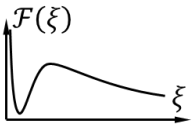
Wenn wir $\mathcal{P}(E, T_1)$ kennen, können wir durch *Neugewichtung* auch $\mathcal{P}(E, T_2)$ vorhersagen (siehe 6.7, NR1):

$$\mathcal{P}(E, T_2) = \frac{\mathcal{P}(E, T_1) e^{-E(\beta_2 - \beta_1)}}{\int dE\, \mathcal{P}(E, T_1) e^{-E(\beta_2 - \beta_1)}}.$$

Wir brauchen $\mathcal{P}$ also im Prinzip nur für eine einzige Temperatur simulieren und können mit obiger Formel auf alle anderen schließen.

Das praktische Problem ist, dass $\mathcal{P}(E)$ in der Regel einen scharfen Peak hat, dessen Lage von der Temperatur abhängt, und die vom Peak entfernten Bins wegen geringer Statistik nur ungenau bekannt sind. Wenn der Peak von $\mathcal{P}(E, T_2)$ dort liegt, wo $\mathcal{P}(E, T_1)$ sehr klein ist, werden die Fehler der obigen Vorhersage-Formel erheblich hochskaliert.

6.3 Potential of Mean Force (PMF)



WAS IST EINE PMF?

Oft möchte man die Freie Energie nur entlang eines Ordnungsparameters ξ berechnen. Die Freie Energie entlang eines Parameters ξ wird als *Potential oft he Mean Force* $\mathcal{F}(\xi)$ bezeichnet. ξ könnte etwa ein Winkel oder Dihedralwinkel sein oder ein Abstand, etwa von einem gelösten Protein zu einer Oberfläche, an der es haftet (die Skizze bezieht sich auf letzteres Beispiel).

FORMALE DEFINITION:

Wir können eine Potential Mean Force $\mathcal{F}(\xi)$ definieren durch

$$\mathcal{F}(\xi) \coloneqq -kT \ln Z(\xi)\,, \qquad Z(\xi) \coloneqq \int d\Gamma\, e^{-\beta E(\Gamma)} \delta(\xi - \xi_\Gamma),$$

wobei ξ_Γ dem Wert des Ordnungsparameters ξ für eine gegebene Konfiguration Γ ergibt. Offenbar gilt $Z = \int d\xi\, Z(\xi)$. $\mathcal{F}(\xi)$ erfüllt die Identität (siehe 6.7, NR2)

$$\int d\xi\, e^{-\beta \mathcal{F}(\xi)} = e^{-\beta F}$$

mit der gesamten Freien Energie F aus 6.1. Die Ableitung von $\mathcal{F}$ ergibt den das thermodynamische Mittel der Kraft $f_\xi \coloneqq -dE/d\xi$ entlang des Weges ξ (siehe 6.7, NR3):

$$\frac{d\mathcal{F}(\xi)}{d\xi} = -\langle f_\xi \rangle.$$

Daher der Name *Potential of the Mean Force*.

ZUSAMMENHANG MIT HISTOGRAMMEN:

Analog zu 6.2 können wir (siehe 6.7, NR4)

$$\mathcal{P}(\xi) = \frac{e^{-\beta \mathcal{F}(\xi)}}{Z} \quad \Leftrightarrow \quad \mathcal{F}(\xi) = -kT \ln \mathcal{P}(\xi) + \text{const}$$

schreiben (ohne Zustandsdichte, da ξ direkt Zustände beschreibt; d. h. $g(\xi) = 1$; außerdem $\ln Z = \text{const}$ bzgl. ξ).

Das Problem ist auch hier wieder: Nehmen wir an, dass sich $\mathcal{F}$ für zwei Parameter ξ_1, ξ_2 nur um $\Delta\mathcal{F} = 4kT$ unterscheidet, so folgt bereits

$$\frac{\mathcal{P}(\xi_2)}{\mathcal{P}(\xi_1)} = e^{-\beta \cdot 4kT} \approx 0{,}02.$$

Nur wenige Einträge landen also im Bin für ξ_2; die Statistik ist sehr schlecht. Eine Lösung dafür ist das Umbrella-Sampling 6.4.

6.4 Umbrella-Sampling

Beim Umbrella-Sampling verschieben wir die Energie E künstlich, sodass Zustände mit relativ niedrigen Energien in den Simulationen dennoch ausreichend Statistik erhalten:

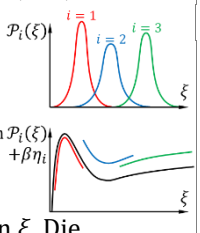
$$E(\Gamma) \to E_i(\Gamma) \coloneqq E(\Gamma) + \eta_i(\xi), \qquad \eta_i(\xi) = \frac{k}{2}(\xi - \xi_i)^2,$$

wobei die Parabel für η_i die gängigste Form ist. Man wählt nun beispielsweise äquidistante ξ_i und lässt für jedes i eine eigene Simulation laufen. Dadurch erhält je stets der Bereich um ξ_i die beste Statistik, auch wenn $E(\xi)$ dort eigentlich groß ist. Was die Verteilungen angeht, so wird

$$\mathcal{P}(\Gamma) \sim e^{-\beta E(\Gamma)} \quad \to \quad \mathcal{P}_i(\Gamma) \sim e^{-\beta E(\Gamma) - \beta \eta_i(\xi)} \sim \mathcal{P}(\Gamma) e^{-\beta \eta_i(\xi)}.$$

Integrieren über Γ liefert mit $\mathcal{P}(\xi) = \int d\Gamma\, \mathcal{P}(\Gamma)\, \delta(\xi - \xi_\Gamma)$

$$\mathcal{P}(\xi) \sim \mathcal{P}_i(\xi) e^{\beta \eta_i(\xi)}.$$



Wenn wir $\mathcal{P}(\xi) \sim e^{-\beta \mathcal{F}(\xi)}$ einsetzen und logarithmieren, folgt

$$-\beta \mathcal{F}(\xi) = \ln \mathcal{P}_i(\xi) + \beta \eta_i(\xi) + \text{const}.$$

Jede Simulation i ergibt – im Prinzip – dasselbe $\mathcal{P}(\xi)$ bzw. $\mathcal{F}(\xi)$, allerdings mit unterschiedlich guter Statistik für unterschiedliche Bereichen von ξ. Die einzelnen Schnipsel von $\mathcal{F}(\xi)$ können dann durch die Wahl der Konstanten verbunden werden. Dazu ist eine hinreichende Überlappung der η_i nötig.

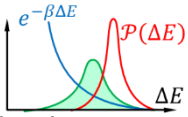
6.5 Freie Energie/Chemisches Potential via Störungen

DIFFERENZ DER FREIEN ENERGIE:

Man betrachte eine Störung, die die Energie von $E_0(\Gamma)$ auf $E_1(\Gamma)$ ändert. Für die freien Energien folgt (siehe 6.1)

$$-\beta\Delta F = \ln\frac{\int d\Gamma\, e^{-\beta\Delta E(\Gamma)} e^{-\beta E_0(\Gamma)}}{\int d\Gamma\, e^{-\beta E_0(\Gamma)} = Z_0} = \ln\int d\Gamma\, e^{-\beta\Delta E(\Gamma)}\mathcal{P}_0(\Gamma)$$
$$= \ln\langle e^{-\beta\Delta E(\Gamma)}\rangle_0 = -\ln\langle e^{\beta\Delta E(\Gamma)}\rangle_1,$$

wobei im zweiten Schritt die Formel für $\mathcal{P}(\Gamma)$ aus 6.2 eingesetzt wurde. Die Formel für den Mittelwert über den 1-Zustand erhält man, indem man $e^{\beta\Delta E(\Gamma)}$ statt in den Zähler in den Nenner einfügt.



Das Problem dieser Formel ist mal wieder die Statistik: Schreiben wir die Formel als $\Delta F \sim \ln\int d\Delta E\, e^{-\beta\Delta E}\mathcal{P}(\Delta E)$, so hat $\mathcal{P}(\Delta E)$ typischerweise einen scharfen Peak, sodass ΔF der grünen Fläche entspricht, die aber zu großen Teilen in Bereichen mit sehr geringen $\mathcal{P}$ liegt, die von der Simulation fast nie erreicht werden. Diese Methode funktioniert nur gut, wenn $\mathcal{P}$ bei kleinen ΔE peakt.

DAS CHEMISCHE POTENTIAL – WIDOM TEST-TEILCHEN:

Das Chemische Potential können wir schreiben als

$$\beta\mu = \beta\frac{\partial F}{\partial N} = \beta\big(F(N+1) - F(N)\big) = \ln\frac{Z(N)}{Z(N+1)}$$
$$= \ln\frac{\int d\Gamma_N\, e^{-\beta E(\Gamma_N)}}{\int d\Gamma_{N+1}\, e^{-\beta E(\Gamma_{N+1})}} = \ln\frac{\int d\Gamma_{N+1}\, e^{-\beta E(\Gamma_N)}}{\int d\Gamma_{N+1}\, e^{-\beta E(\Gamma_{N+1})}}.$$

Zuletzt haben wir angenommen, dass das Integrationsmaß des Zustandsraums so gewählt ist, dass $\int d\Gamma = 1$, sodass wir die Integration über das $N+1$-te Teilchen in das Integral oben integrieren können, ohne etwas zu verändern. Wir können als Störung nun das „Einschalten" der Wechselwirkungen dieses $N+1$-sten Teilchens betrachten:

$$E_0(\Gamma_{N+1}) \coloneqq E(\Gamma_N), \qquad E_1(\Gamma_{N+1}) \coloneqq E(\Gamma_{N+1}).$$

Damit folgt mit gleicher Rechnung wie bei der Freien Energie

$$\beta\mu = -\ln\frac{\int d\Gamma_{N+1}\, e^{-\beta E_1(\Gamma_{N+1})}}{\int d\Gamma_{N+1}\, e^{-\beta E_0(\Gamma_{N+1})}} = -\ln\langle e^{-\beta\Delta E(\Gamma_{N+1})}\rangle_0.$$

Um μ zu bestimmen, können wir in Simulationen also in jedem Simulationsschritt ein Testteilchen einfügen, ΔE berechnen, das Testteilchen wieder entfernen und nach dem nächsten Schritt wieder einfügen. Mit all den daraus gewonnenen ΔE-Werten können wir obige Mittelung durchführen. Statt ein Testteilchen hinzuzufügen, kann man auch eines entfernen; mit analoger Herleitung findet man dann $\beta\mu = \ln\langle e^{\beta\Delta E}\rangle_1$.

6.6 Thermodynamische Integration

EINFACHE THERMODYNAMISCHE INTEGRATION:

Ableiten der Freien Energie nach dem Volumen ergibt den Druck

$$P = -\frac{\partial F}{\partial V}.$$

Wenn wir also die Funktion $P(V)$ berechnen könnten, könnten wir Differenzen Freier Energien berechnen durch

$$\Delta F = F(V_1) - F(V_0) = -\int_{V_0}^{V_1} dV\, P(V).$$

Wir könnten also Simulationen bei unterschiedlichen Volumen zwischen V_0 und V_1 laufen lassen, um $P(V)$ zu finden und daraus ΔF zu berechnen.

INTEGRATION ENTLANG EINES PARAMETERS:

Betrachten wir eine Energiefunktion $E(\Gamma, \lambda)$, die von irgendeinem Parameter λ abhängt. Dann gilt (siehe 6.7 NR5)

$$\frac{dF}{d\lambda} = \left\langle\frac{dE}{d\lambda}\right\rangle_\lambda \quad\Longrightarrow\quad F(\lambda_1) - F(\lambda_0) = \int_{\lambda_0}^{\lambda_1} d\lambda\, \left\langle\frac{dE}{d\lambda}\right\rangle_\lambda.$$

Der Index λ drückt aus, dass der Mittelwert für ein bestimmtes λ zu bilden ist, sodass $\langle dE/d\lambda\rangle_\lambda$ noch von λ abhängt. Die Simulation muss also über verschiedene λ zwischen λ_0 und λ_1 laufen, um daraus den Mittelwert von $dE/d\lambda$ zu bilden.

6.7 Nebenrechnungen

NEBENRECHNUNG 1:

Wir verwenden die Formeln für $\mathcal{P}, \mathcal{S}$ und F und erhalten

$$\mathcal{P}(E, T_2) = \frac{e^{\mathcal{S}(E)-\beta_2 E}}{Z(T_2)} = \frac{e^{(\ln\mathcal{P}(E,T_1)+\beta_1 E-\beta_1 F)-\beta_2 E}}{Z(T_2)}$$
$$= \mathcal{P}(E, T_1) e^{-E(\beta_2-\beta_1)}\frac{Z(T_1)}{Z(T_2)}.$$

$Z(T_1)/Z(T_2)$ ist von E unabhängig. Integrieren wir obige Gleichung über E und nutzen $\int dE\, \mathcal{P} = 1$, folgt

$$\frac{Z(T_1)}{Z(T_2)} = \frac{1}{\int dE\, \mathcal{P}(E, T_1) e^{-E(\beta_2-\beta_1)}}.$$

Dies können wir wiederum einsetzen, um die gewünschte Gleichung zu erhalten.

NEBENRECHNUNG 2:

$$\int d\xi\, e^{-\beta\mathcal{F}(\xi)} = \int d\xi \int d\Gamma\, e^{-\beta E(\Gamma)}\delta(\xi - \xi_\Gamma) = \underbrace{\int d\Gamma\, e^{-\beta E(\Gamma)}}_{=Z}$$
$$= e^{\beta kT \ln\int d\Gamma\, e^{-\beta E(\Gamma)}} = e^{-\beta F}$$

NEBENRECHNUNG 3:

Mit der mathematischen Identität

$$\frac{d}{da}\int dx\, g(x)\, \delta(x-a) = \int dx \frac{dg(x)}{dx}\delta(x-a)$$

folgt direkt durch mehrfache Anwendung der Kettenregel

$$\frac{d\mathcal{F}(\xi)}{d\xi} = -\frac{\int d\Gamma\, f_\xi e^{-\beta E(\Gamma)}\delta(\xi-\xi_\Gamma)}{\int d\Gamma\, e^{-\beta E(\Gamma)}\delta(\xi-\xi_\Gamma)} = -\langle f_\xi\rangle, \quad f_\xi = -\frac{dE(\Gamma)}{d\xi}.$$

Unter dem Integral kann man wegen der δ-Funktion jederzeit $d/d\xi_\Gamma \to d/d\xi$ ersetzen.

NEBENRECHNUNG 4:

Es gilt mit $\mathcal{P}(\Gamma) = e^{-\beta E(\Gamma)}/Z$

$$\frac{e^{-\beta\mathcal{F}(\xi)}}{Z} = \frac{Z(\xi)}{Z} = \frac{\int d\Gamma\, e^{-\beta E(\Gamma)}\delta(\xi-\xi_\Gamma)}{Z} = \frac{\int d\Gamma\, Z\mathcal{P}(\Gamma)\,\delta(\xi-\xi_\Gamma)}{Z}$$
$$= \int d\Gamma\, \mathcal{P}(\Gamma)\,\delta(\xi-\xi_\Gamma) = \mathcal{P}(\xi).$$

NEBENRECHNUNG 5:

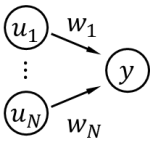
$$\frac{dF}{d\lambda} = -kT\frac{d}{d\lambda}\ln\int d\Gamma\, e^{-\beta E(\Gamma,\lambda)} = -kT\frac{-\beta\int d\Gamma \frac{dE}{d\lambda} e^{-\beta E(\Gamma,\lambda)}}{\int d\Gamma\, e^{-\beta E(\Gamma,\lambda)}}$$
$$= \left\langle\frac{dE}{d\lambda}\right\rangle_\lambda.$$

7 Neuronale Netze

7.1 Das Perceptron

AUFBAU:

Das Perceptron ist die einfachste Form eines neuronalen Netzwerks, bestehend aus nur einem Neuron. Der Input an den N Dendriten kann als Vektor $\vec{u} = (1, u_1, \dots, u_N)$ geschrieben werden die mit Gewichten $\vec{w} = (w_0, w_1, \dots, w_N)$ versehen werden. Der Input des Neurons insgesamt ist dann das Skalarprodukt $\vec{u} \cdot \vec{w}$. Man beachte, dass es zusätzlich zu den Dendriten ein Offset w_0 gibt.



Alle Dendriten hängen an einem Neuron, das den Wert

$$y = f(\vec{u} \cdot \vec{w})$$

annimmt. $f: \mathbb{R} \to Z$ ist dabei eine Bewertungsfunktion; in der Regel ist $Z = [0, 1]$ oder $[-1, 1]$. Ein sehr einfacher Fall wäre $f(x) = \theta(x)$ mit der Heavy-Side-Funktion, aber auch $f(x) = (1 + e^{-ax})^{-1}, a \in \mathbb{R}$ ist sicherlich oft sinnvoll.

FUNKTION:

Das Perceptron soll lernen, für bestimmte Input-Daten $\vec{u}$ bestimmte Output-Daten y zu liefern. „Lernen“ bedeutet, die korrekten Werte für die Komponenten von $\vec{w}$ zu finden, sodass er zu beliebigen Input-Daten $\vec{u}$ den korrekten Output y findet.

FLIESEN-BEISPIEL:

Bei der Herstellung von Fliesen können unsichtbare Risse entstehen. Geübte Arbeiter erkennen beschädigte Fliesen an dem Ton, den sie erzeugen, wenn sie mit einem Hammer draufschlagen. Diesen Job soll nun ein Perceptron lernen. Das Geräusch wird aufgenommen und fourierzerlegt; $\vec{u}$ enthält dann die Amplituden verschiedener Frequenzen. Es gibt nun einen (im folgenden beschriebenen) Algorithmus, mit dem man die w_i so finden kann, dass $y = 0$ für defekte und $y = 1$ für intakte Fliesen gilt. Der Algorithmus wird dabei nur mit korrekten Beispiel-Tupeln $(\vec{u}, y_0)$ gespeist, die der Arbeiter liefern kann.

LERN-ALGORITHMUS:

Das Perceptron lernt an bekannten Beispielen $(\vec{u}, y_0)$. Wir starten mit einem zufälligen Vektor $\vec{w}$ und füttern das Perceptron mit einem beispielhaften $\vec{u}$, von dem wir das richtige Ergebnis y_0 schon kennen. Falls

$$y = f(\vec{u} \cdot \vec{w}) = y_0,$$

lassen wir $\vec{w}$ unverändert. Anderenfalls gibt es einen Fehler

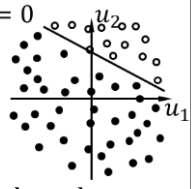
$$\begin{aligned} \delta = y - y_0 = f(\vec{u} \cdot \vec{w}) - f(\vec{u} \cdot \vec{w}_0) &\approx (\vec{w} - \vec{w}_0) \cdot \nabla_{\vec{w}} f(\vec{u} \cdot \vec{w}) \\ \Leftrightarrow \quad \vec{w}_0 \cdot \nabla_{\vec{w}} f(\vec{u} \cdot \vec{w}) &\approx \vec{w} \cdot \nabla_{\vec{w}} f(\vec{u} \cdot \vec{w}) - \delta \\ \Leftrightarrow \quad \vec{w}_0 &\approx \vec{w} - \delta \frac{\nabla_{\vec{w}} f(\vec{u} \cdot \vec{w})}{\left(\nabla_{\vec{w}} f(\vec{u} \cdot \vec{w})\right)^2}. \end{aligned}$$

Da diese Gleichung nur ungefähr gilt ($\approx$), müssen wir uns dem Endergebnis iterativ annähern. Am Ende jedes Durchlaufs korrigiert man daher $\vec{w} \to \vec{w}'$ mit

$$\vec{w}' = \vec{w} - \eta \delta \, f'(\vec{u} \cdot \vec{w}) \, \vec{u}.$$

Der Gradient im Nenner wird zugunsten eines kleinen Parameters η vernachlässigt. Mit der Kettenregel erhält man $\nabla_{\vec{w}} f = \vec{u} f'$. Nun kann das System in jeder Iteration mit neuen bekannten Paaren $(\vec{u}, y_0)$ gefüttert werden, um $\vec{w}$ zu „erlernen“. ($f' = 1$ für $f = \theta$)

VERANSCHAULICHUNG:

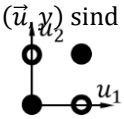


Alle Input-Daten $\vec{u}$ kann man als Punkte in einem $\vec{u}$-Raum verstehen. Für ein gegebenes $\vec{w}$ definiert $\vec{u} \cdot \vec{w} = 0$ eine bestimmte Ebene in diesem Raum. Man betrachte $f = \theta$; dann liefern alle Punkte diesseits der Ebene $y = 0$ und jenseits $y = 1$. Eine korrekte Bestimmung von $\vec{w}$ entspricht also der Positionierung der Ebene $\vec{u} \cdot \vec{w} = 0$ derart, dass alle „richtigen“ Punkte auf der einen und alle „falschen“ auf der anderen Seite liegen. Man beachte, dass wegen $u_0 = 1$ die Ebenen einen Ursprungsabstand $w_0 > 0$ haben können.

7.2 Multilayer-Perceptron: XOR

XOR-PROBLEM DES PERCEPTRONS:

Aus der Veranschaulichung in **Fehler! Verweisquelle konnte nicht gefunden werden.** wird offensichtlich, dass das einfache Perceptron nur für solche Anwendungen funktionieren kann, wenn die beiden Punktarten im $\vec{u}$-Raum durch eine Ebene trennbar sind. Das einfachste Beispiel, was sich damit nicht mehr lösen lässt, ist ein 2D-exclusive-or-Problem (XOR). Hierbei ist $\vec{u} = (1, u_1, u_2)$ mit $u_i \in \{0, 1\}$ und sämtliche Tupel $(\vec{u}, y)$ sind bekannt/gegeben als

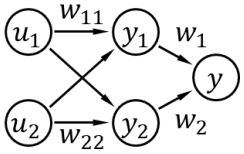


$$y = u_1 \text{ XOR } u_2.$$

Hier lassen sich die $y = 0$-Punkte nicht von den $y = 1$-Punkten durch eine Ebene separieren.

MULTILAYER-PERCEPTRON FÜR XOR:

Allerdings lässt sich das Problem lösen, in dem man eine Zwischenebene einbaut:



Jedes Neuron (y_1, y_2, y) hat hier einen Gewichte-Vektor mit drei Komponenten:

$$\begin{aligned} y_1\text{:} \quad & \vec{w}_1 = (w_{10}, w_{11}, w_{12}) = (-1/2\,, 1, -1), \\ y_2\text{:} \quad & \vec{w}_2 = (w_{20}, w_{21}, w_{22}) = (1/2\,, 1, -1), \\ y\text{:} \quad & \vec{w} = (w_0, w_1, w_2) = (1/2\,, 1, -1). \end{aligned}$$

(w_{12}, w_{21} mangels Platzes nicht beschriftet). Die eingetragenen Werte für die Gewichte beweisen, dass dieses Perceptron XOR für $f = \theta$ lösen kann; mit $\vec{y} \coloneqq (1, y_1, y_2)$ folgt

$$\begin{aligned} y_i = \theta(\vec{u} \cdot \vec{w}_i) &= \theta(\mp 1/2 + u_1 - u_2) \\ \Rightarrow \quad y = \theta(\vec{y} \cdot \vec{w}) &= \theta(1/2 + y_1 - y_2) \\ &= \theta\big(1/2 + \theta(-1/2 + u_1 - u_2) - \theta(1/2 + u_1 - u_2)\big). \end{aligned}$$

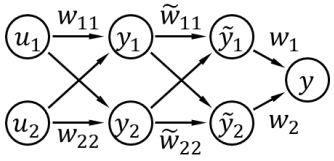
Für die verschiedenen Input-Vektoren $\vec{u}$ erhalten wir somit

$$y = \begin{cases} \theta(1/2 + 0 - 1) = 0, & \text{für } u_1 = u_2, \\ \theta(1/2 + 1 - 1) = 1, & \text{für } u_1 = 1, u_2 = 0, \\ \theta(1/2 + 0 - 0) = 1, & \text{für } u_1 = 0, u_2 = 1. \end{cases}$$

7.3 Back-Propagation beim Multilayer-Perceptron

Wie funktioniert nun der Lern-Algorithmus beim Multilayer-Perceptron?

Füttert man das Perceptron wieder mit bekannten Beispielen $(\vec{u}, y_0)$, kann man am Ende wieder einen Fehler $\delta = y - y_0$ berechnen. Kanten mit größerem momentanem Gewicht werden zu diesem Fehler verstärkt beigetragen haben, daher proagieren wir den Fehler gewichtet zurück; beispielsweise ordnen wir jedem Neuron des Netzes



folgende Fehler zu:

$$\delta = y - y_0, \qquad \tilde{\delta}_i = w_i \delta, \qquad \delta_i = \Sigma_j \tilde{w}_{ij} \tilde{\delta}_j.$$

Wie in **Fehler! Verweisquelle konnte nicht gefunden werden.** können wir nun wieder die Gewichte anpassen:

$$\begin{aligned} w'_{ij} &= w_{ij} - \eta \delta_j \, f'_j(\vec{u} \cdot \vec{w}_i) \, u_i, \\ \tilde{w}'_{ij} &= \tilde{w}_{ij} - \eta \tilde{\delta}_j \, \tilde{f}'_j\big(\vec{y} \cdot \vec{\tilde{w}}_i\big) \, y_i, \\ w'_i &= w_i - \eta \delta \, f\big(\vec{\tilde{y}} \cdot \vec{w}\big) \, \tilde{y}_i. \end{aligned}$$