

experimentelle Methoden der Bioinformatik

Inhaltsverzeichnis

1	Allgemein / Hintergrund	1
2	ChIP-Chip und ChIP-Seq	1
2.1	Ablauf	1
2.1.1	Crosslinking	1
2.1.2	Sonication	1
2.1.3	Immunoprecipitation (Selektion mittels Antikörper)	1
2.1.4	Reverse Immunoprecipitation	2
2.1.5	Reverse Cross Linking	2
2.1.6	Auswertung	2
2.2	Probleme/Fehler	2
2.3	Antikörper	3
3	Peak Calling	4
3.1	MACS	5
4	CLIP-Seq	6
4.1	ICLIP	6
5	PAR-CLIP	6
6	Protein-Protein-Interaktion	6
7	Tandem Affinity Purification (TAP)	6
7.1	Local clique merging algorithm (LCMA)	6
7.2	Clique Finding Algorithm (CFA)	6
8	RNA structure probing	8
8.1	objective function approach	8
8.2	Inline-Probing	8
8.3	Chemisches Probing	9
8.3.1	MACS (Model-based Analysis of ChIP-Seq)	9
8.3.2	CLIP: Cross-linking and immunoprecipitation protocol . .	9
8.3.3	SHAPE-Seq	9
8.3.4	Hydroxyl-Radikal Probing	11
8.3.5	DMS	11
8.3.6	CMCT	11
8.3.7	Kethoxal	11
8.4	Nucleotide analoge interference mapping (NAIM)	12
9	Proteinstrukturen	13
10	X-ray crystallography	14

1 Allgemein / Hintergrund

Messung von Strukturen vs. Messung von Interaktionen

Motifsuche:

- Proteine (Transkriptionsfaktoren) haben Domäne die Nukleotidsequenzen erkennen
- Position weight matrix (PWM), position specific scoring matrix (PSSM)
- MEME zum erkennen von Sequenzen / Motifen

2 ChIP-Chip und ChIP-Seq

ChIP: **Ch**romatin-**I**mmuno**P**recipitation

Kein Single Cell Protocol -> es werden Zellpopulationen benötigt

Ziel: Man will feststellen an welcher Stelle Proteine binden

Quellen für Fehler / Ungenauigkeiten: Messung des Populationsmittelwerts

ChIP-Chip: Chromatin-Immunoprecipitation Chip

ChIP-Seq: Chromatin-Immunoprecipitation DNA-Sequencing

2.1 Ablauf

2.1.1 Crosslinking

Stabilisierung der Bindungen zwischen DNA und Protein

Geschieht reversibel zwischen DNA (**Chromatin**) und rekombinanten Proteinen

- Formaldehyd (CH₂O) vernetzt Base (B) mit Proteinen (P-NH₂) quer
- $\text{P-NH}_2 + \text{CH}_2\text{O} \rightleftharpoons \text{PN}=\text{CH}_2 + \text{NH}_2\text{-B} \rightleftharpoons \text{PNH-CH}_2\text{-NH-B}$
- Rekombinant: Biotechnologisch hergestellte Proteine aus genetisch veränderten Organismen

2.1.2 Sonication

Zerstören und Zerkleinern (fragmentieren) der Zellen, Zellbestandteile und DNA durch Ultraschall

(Vorher: Waschen der Zellen mit Protease Inhibitor, Lyse + homogenisieren)

- zeitkritisch → Länge bestimmt Grad der Zerkleinerung
- 200-1000 BP Fragmente im Idealfall

Ergebnis sind DNA Fragmente mit gebundenen Proteinen

2.1.3 Immunoprecipitation (Selektion mittels Antikörper)

- Antikörper (binden an Beads oder Membranen, Chip/in Gel) binden an rekombinante Proteine

oder Protein-TAG (kurze Aminosäuresequenz, markieren Protein)

- Aufreinigung:

- Zentrifugation des Präzipitats: Beads+(Protein-DNA) am Boden, Zellfragmente/Rest in Lösung
- Abkippen der Lösung
- Aufnehmen des Beadspellets in Puffer, erneut zentrifugieren (x-Mal)
- Manchmal noch
- DNase Verdau der DNA in Lösung
- Aufheben der DNA in Lösung, als total-Chromatin-Probe

2.1.4 Reverse Immunoprecipitation

Durch Aufreinigungsschritte sind Beads/Gel/Chip idealerweise frei von Zellfragmenten/ungebundener DNA.

Umkehren der IP mit Elutionspuffer → Antikörper von DNA+Proteine trennen
→ Salzgehalt und PH-Wert an Rückreaktion angepasst

2.1.5 Reverse Cross Linking

- Thermische Zerstörung der Bindung zw. Protein und DNA
- Salzgehalt des Buffer angepasst auf Rückreaktion - Proteinase K und RNase bauen Proteine und RNA ab (zur Aufreinigung)
- Extraktion der übrig gebliebenen DNA durch Zentrifuge

2.1.6 Auswertung

Chiphybridisierung

- Hybridisierung der DNA an Microarray
- Färbung der DNA
- Messung der Farbintensität

→ *mit dem ChIP Background kann ich nichts anfangen...* ←

Sequencing

Hochdurchsatzsequenzierung der aufgereinigten DNA.

- DNA extrahieren → DNA fragmentieren → Primer an Fragmente → Sequenzierung
- Herausrechnen der Primer (idealerweise kennt man sie) →
- Quality control → Phred-score Berechnung (Güte der erkannten Nukleobase) → Cutoff bei zu niedrigem Phred-score → Mapping des sequenzierten Teilstücks auf Genom

2.2 Probleme/Fehler

Cross-Linking

FN: Protein an DNA gebunden, aber kein Cross-Linking

FP: Proteine, die sehr nahe an der DNA sind, aber ungebunden, werden

auch cross linked

Sonication

- Größe der Fragmente abhängig von Ultraschalleinsatz – zeitkritisch!
- Kürzere und längere Fragmente können Informationen enthalten

Immunoprecipitation

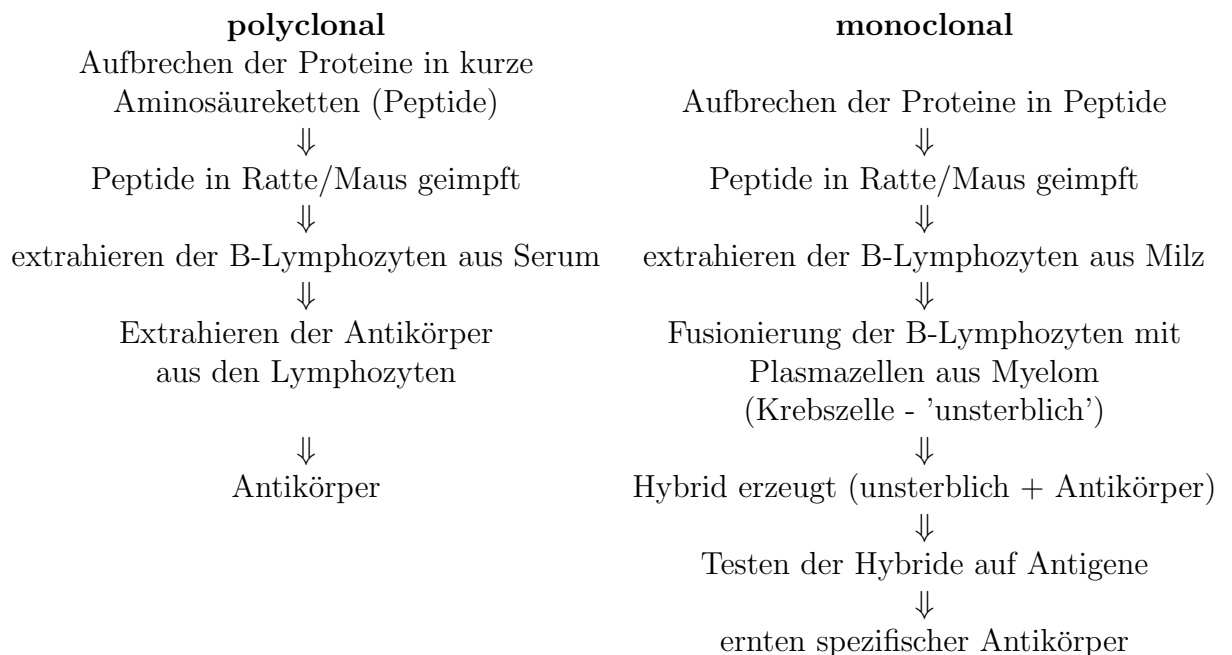
FP: Mangelnde Reinheit der rekombinanten Proteine; Spezifität der heterophilen Antikörper zu gering
Aufreinigung führt zu **FP** und **FN**

Chip

FN: Hybridisierung nicht effektiv genug

2.3 Antikörper

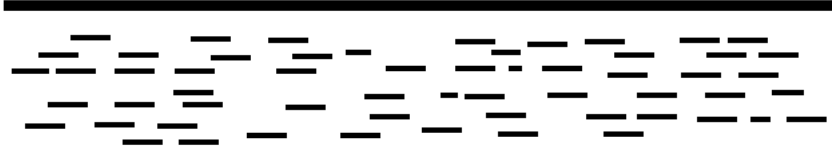
- Antikörper bindet spezifisch und sensitiv
- Antikörper sind fixiert an:
 - Beads
 - Chip (kein Microarray)
 - Gel
- Antikörper werden im Experiment erzeugt



3 Peak Calling

Sequenziertes Genom/RNA/DNA aus dem Experiment = viele, kurze Reads

→ naiv: Jedes Nukleotid, dass von Reads bedeckt ist = Gebunden



→ Problem: Viele FP, da kurze Reads mehrere Treffer haben können

→ Lösung: Cutoff für Anzahl der Reads auf Nukleotid

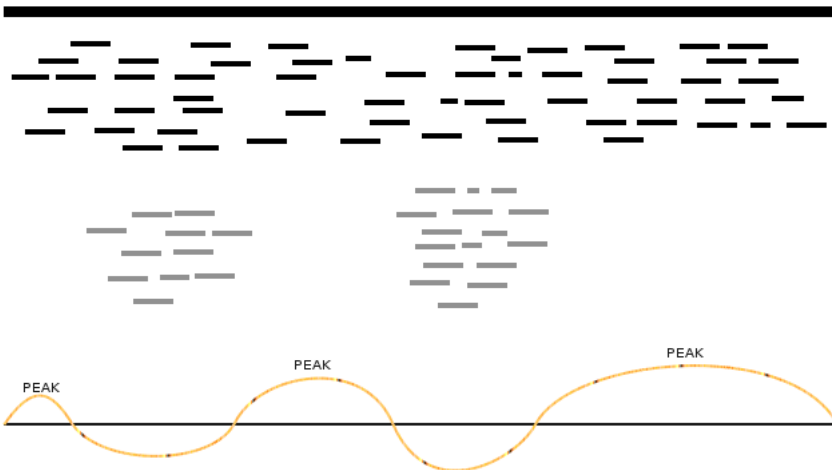
→ Problem: Manche Basen einfach zu binden = viele FP

So geht das nicht!

Lösung:

Enrichment: $\log \frac{Expression}{Background}$

naiv: Wenn Enrichment > Cutoff → Peak!



3.1 MACS

Model-based Analysis of ChIP-Seq (MACS)

1. Einteilen des Genoms in Bins (Eimer)

Window: 200 BP und Offset von 1/4 der window size

In Bins werden Reads eingeordnet

2. Zählen der Fragmente pro Bin, +/- Strang

→ Poisson verteilt!

$$P(x > k, \lambda) = \sum_{i=k}^{\infty} P\lambda(i) = 1 - \sum_{i=0}^{k-1} P\lambda(i) = 1 - \sum_{n=0}^{k-1} \frac{\lambda^n}{n!} e^{-\lambda}$$

λ =Mittelwert der read counts aus Background, k =read counts aus Experiment
read count signifikant größer Mittelwert → Peak!

Mittelwert kann abhängig von Menge der reads in Window sein:

$$\lambda = \max(\lambda_{\text{global}}, \lambda_{1000}, \lambda_{5000}, \lambda_{10000})$$

→ Window jeweils zentriert an Bin

3. p-Value Correction

Holm-Bonferroni

q-Value

4. Peakmerging

Wenn Abstand zwischen Peaks < Cutoff → Merge Peaks

(bei MACS 2xWindowSize)

Wo sind die Bindungsstellen?

Protein → RNA - **ChIP**: Regionen, mit denen das Protein assoziiert ist

DNA → RNA - **ChIRP**: Match von RNA auf sequenzierter DNA

Verfahren ähnlich zu CLIP

→ RNA cross-linking (UV o. formalin) → aufreinigen →

reverse cross-linking → Read → Match mit DNA

(Chromatin isolation by RNA purification)

Protein → RNA - **RIP**: RNA zu cDNA, hybridisieren mit Chip

→ RNA cross-linking (UV o. formalin) → aufreinigen →

reverse cross-linking → RNA in cDNA →

Hybridisierung auf Chip

(RNA immunoprecipitation protocol)

4 CLIP-Seq

4.1 ICLIP

5 PAR-CLIP

6 Protein-Protein-Interaktion

7 Tandem Affinity Purification (TAP)

7.1 Local clique merging algorithm (LCMA)

7.2 Clique Finding Algorithm (CFA)

MACS (Model-based Analysis of ChIP-Seq):

1) Einteilen des Genoms in Bins

Window Size: typisch 200 bp & offset (ungefähr 0,25 windows size = 50 bp)

MACS empfiehlt Bin doppelt so groß wie Fragment

2) Zähle die Anzahl an hypothetischen Fragmenten pro Bin (=window)

Fragmente können in mehr als ein Bin fallen

CLIP Cross-linking & immunoprecipitation protocol - Ultraviolettes Licht für cross linking - UV cross linked nur RNA mit Proteinen - induziert UV Mutation der RNA - CIMS: cross-linking induced mutation sites

8 RNA structure probing

Bestimmung von:

- Basenpaarung
- Sekundärstruktur und Tertiärstruktur

8.1 objective function approach

Hard constraints:

→ 3 Aussagen möglich: — = gepaart; . = ungepaart; X = unbekannt

Soft constraints:

→ Wahrscheinlichkeit ob Base an Position Y gepaart ist oder nicht

→ Minimiere den Fehler $F(\vec{E})$

$$\vec{E} = \sum_{\mu} \frac{\varepsilon_{\mu}^2}{\tau^2} + \sum_{i=0}^n \frac{1}{\sigma^2} (p_i(\vec{\varepsilon}) - q_i)^2 \quad (1)$$

μ ... Strukturelemente ε_{μ} ... Betrag der Stör-Energie eines Strukturelements

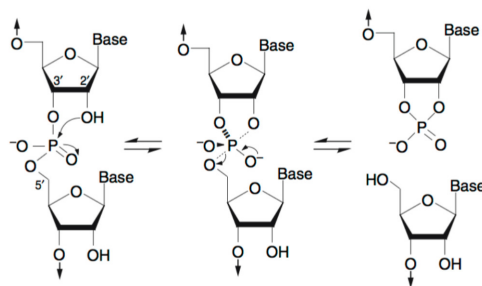
τ^2 ... Varianz des Standardenergiemodells

σ^2 ... Varianz der Probingdaten

$p_i(\vec{\varepsilon})$... Wahrscheinlichkeit, dass i ungepaart ist unter Bedingung des Standardenergiemodells und der Störenergie

8.2 Inline-Probing

inline-nucleophilic-attack: Wie in der Abbildung zu sehen kommt es zu strukturellen Änderungen der chemischen Konformation des RNA-Strangs an der Phosphatgruppe. Grund hierfür ist die Instabilität der Einzelsträngigen RNA, die bei Bindung eines Liganden an das Molekül zum Bruch (Cleavage") führt oder eine rein zufällige Konformationsänderung des RNA-Moleküls.



Vorgehen:

- Erstellen von zwei Proben des zu untersuchenden RNA-Moleküls

- In einer Probe gewählten Ligand hinzugeben
- beide Proben werden lange inkubiert → nucleophilic attack
- Gelbild mittels Gelelektrophorese herstellen und Längen der RNA-Fragmente beider Proben vergleichend betrachten
- gleiche Strukturen werden als Hintergrundrauschen (ligandenunabhängige) Cleavages betrachtet

8.3 Chemisches Probing

RNA-modifizierende Chemikalien sind **struktursensitiv** [1] und **sequenzunabhängig**

- 1 Es werden Chemikalien genutzt die entweder gepaarte oder ungepaarte Basen modifizieren
- 2 Mechanismus zur Detektion der Modifikation

8.3.1 MACS (Model-based Analysis of ChIP-Seq)

- Einteilen des Genoms in Bins
 - a Windowsize: typisch 200 bp + offset (ungefähr 0,25 windowsize = 50 bp)
 - b MACS empfiehlt Bin doppelt so groß wie Fragmente
 - Zähle die Anzahl an hypothetischen Fragmenten pro Bin (=window)
- Fragmente können in mehr als ein Bin fallen

8.3.2 CLIP: Cross-linking and immunoprecipitation protocol

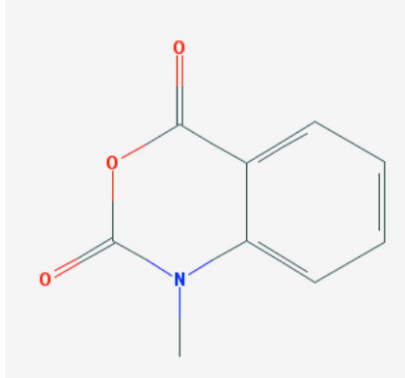
- Ultraviolettes Licht für cross linking
- UV cross linked nur RNA mit Proteinen
- induziert UV Mutation der RNA
- CIMS: cross-linking induced mutation sites

8.3.3 SHAPE-Seq

(Selective 2'-hydroxylacetylation analyzed by **p**rimers **e**xtension **s**equencing)

- 2'-OH ist reaktiver wenn die zugehörige Base ungebunden ist
- genutzte Chemikalie: N-methylisatoic anhydride

- unter Abgabe von Kohlenstoffdioxid (CO_2) bindet ein Sauerstoffmolekül des NMIA an 2'-OH der RNA



(Quelle: https://pubchem.ncbi.nlm.nih.gov/compound/N-Methylisatoic_anhydride)

- reverse Transkription: Die RNA wird mit DNA-Molekülen transkribiert. Im Anschluss werden die gewonnenen DNA-Fragmente sequenziert und als Library gespeichert
- Da es auch zu zufälligen Abbruch bei der reversen Transkription kommen kann, wird ebenfalls eine negativ-Library erzeugt
- Alignment der Reads an Transkriptom der RNA (X_{ij} , wobei i = Basenposition, j = Library)
- Maximum-Likelihood-Model:
 - $r_i = \frac{r_{i+}}{r_{i-}} \rightarrow$ Datengrundlage
 - negativ-Library \rightarrow Abbruchrate
 - simulierte Daten $m_i \rightarrow$ Berechnung der positionsweisen Shape-Reaktivität

\rightarrow Ermittlung der pseudo-Free-Energy

$$\Delta G_{Shape_i} = m * \ln(\gamma_i * 1, 0) + b \quad (2)$$

m ... Anstieg des Bestrafungswertes

$1,0$... Pseudocount b ... negativer Bonus der freien Energie für gepaarte Basen

$$M_{ij} = \min \begin{cases} M(i+1, j) \\ \min(M(i+1, k-1) * M(k+1, j) * e^{-\frac{E'_{ij}}{kT}}) \end{cases} \quad (3)$$

wobei:

$$E'_{ij} = E_{ij} + \Delta G_{Shape_i} + \Delta G_{Shape_j} \quad (4)$$

E_{ij} ... Standard Energiemodell

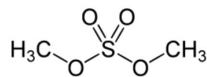
8.3.4 Hydroxyl-Radikal Probing

Hydroxyl-Radikale führen zum Bruch der RNA-Sequenz, wenn keine 3-D Interaktion stattfindet und keine Bindung an ein Protein vorliegt.

Nachteil: Sie sind nur kurzlebig in Lösung und müssen hergestellt werden

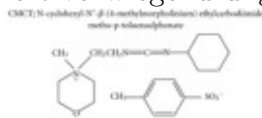
8.3.5 DMS

Di-Methylsulfat bindet an CH_3 von ungebundenen A bzw. C oder an eines der beiden, wenn sie das letzte Basenpaar einer Helix bilden oder wenn sie direkt neben einem GU-Basenpaar liegen.



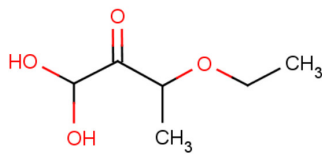
8.3.6 CMCT

(1-Cyclohexyl-(2-Morpholinoethyl)Carbodiimid Metho-p-Toluensulfonat) modifiziert vorwiegend ungepaartes Uridin und teilweise ungepaartes Guanin.

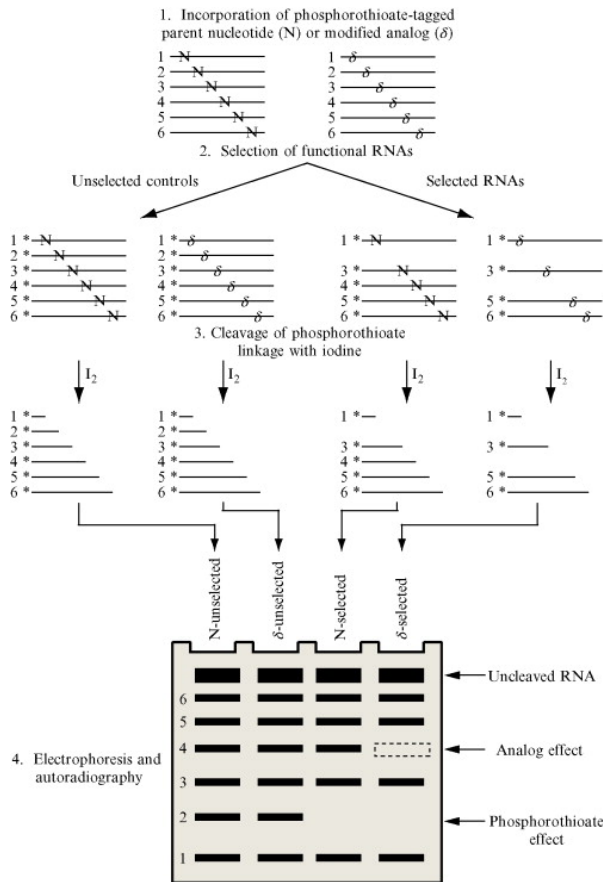


8.3.7 Kethoxal

Kethoxal modifiziert ungepaartes Guanin



8.4 Nucleotide analogue interference mapping (NAIM)



(Quelle: <http://www.sciencedirect.com/science/article/pii/S0076687909680010>)

NAIM ist eine Erweiterung des Interferenz-Mappings mit Triphosphorsäure-Substitution. Untersucht, welche Basen funktional sind. Vorgehensweise:

- Nukleotide sind prinzipiell ohne funktionelle Gruppe
- Nukleotide werden in vitro zufällig durch getaggende Analogika und getaggende normale Nukleotide während Transkription markiert
- Annahme: Jedes Transkript hat nur ein getaggendes Nukleotid/Analogon
- Auswahl der aktiven funktionalen RNAs und Erzeugung einer inaktiven Kontrollgruppe
- Cleavage (Beschneiden) hinter der getaggenden Struktur durch Iod
- Gelelektrophoresebild → gibt Aussage darüber, welche durch Selektion sichtbar werden und welche durch Nukleotid-Einbau sichtbar sind

9 Proteinstrukturen

Methoden

- NMR-Spektroskopie (Protein in Lösung)
- Röntgen-Kristallographie (Protein als Kristall)

→ Bestimmung der 3D-Atompositionen → Position-Database (PDB)

Nachteil: sehr ungenau und starkes Hintergrundrauschen

10 X-ray crystallography

Voraussetzung: regulären Kristall aus dem Protein



Bragg's Law: $n\lambda = 2d\sin(\Theta)$

X-ray crystallography diffraction:

X-ray \rightarrow Kristall \rightarrow Ablenkung

durch Atome \rightarrow Ablenkung wird durch einen Detektor gemessen

fixe Wellenlänge λ , Winkel Θ variieren (Kristall rotieren) \rightarrow charakteristisches

Diffraction pattern \rightarrow Amplitude ändert sich über den Winkel

$d_{hkl} = \frac{a_0}{\sqrt{h^2 + k^2 + l^2}}$ mit hkl =Laue-Index, a_0 = Gitterkonstante

oder:

Θ fest und λ variieren \rightarrow white x-ray



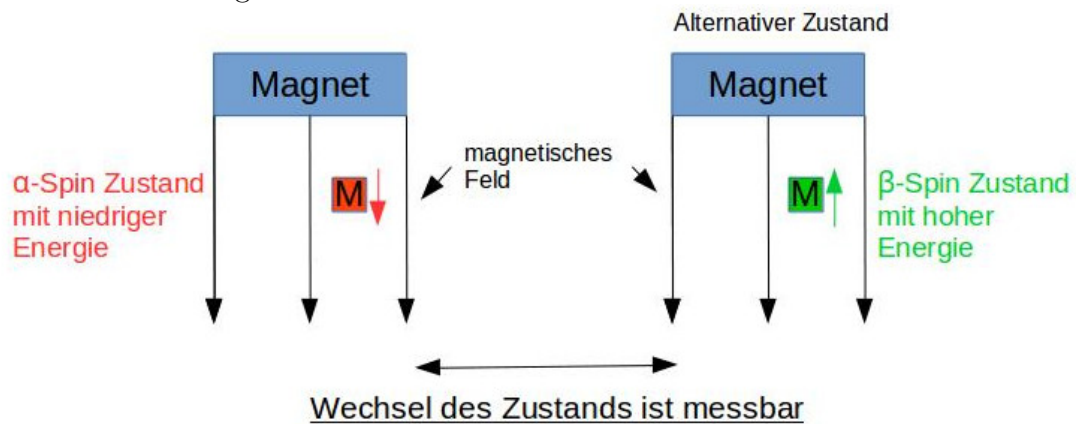
rotationssymmetrisches
Punktmuster

Kombinierte Information aus allen Messungen für verschiedene λ & Θ

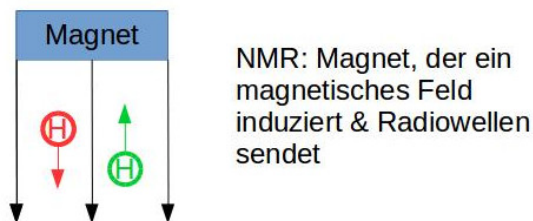
1. Backbone des Proteins ($COOH - NH_2$)
2. Bestimmung der Position der flexiblen Seitenketten der Aminosäuren
3. Verbesserung

11 NMR spectroscopy

NMR: nuclear magnetic resonance



Atome mit magnetischen Eigenschaften: H, Deuterium, N, C, Li, B, O



- ohne weitere äußere Einflüsse Atom in α - spin
- über Flips im Magnetfeld Ermittlung der Protein-Struktur

Spektren von H,C,N + Strukturformel der bekannten Aminosäure + Aminosäureketten
 → Wechselwirkungen zwischen den Gruppen herleiten → 3D Koordinaten berechnen