experimentelle Methoden der Bioinformatik

Inhaltsverzeichnis

1 Allgemein / Hintergrund

Messung von Strukturen vs. Messung von Interaktionen
Motifsuche:
- Proteine (Transkriptionsfaktoren) haben Domaine 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 (**DNA-Protein-Interaktion**[1])
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
- P-NH2+CH2O $\rightleftharpoons$ PN=CH2+NH2-B $\rightleftharpoons$ PNH-CH2-NH-B
- Rekombinant: Biotechnologisch hergestellte Proteine aus genetisch veraenderten 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

[1] `https://de.wikipedia.org/wiki/Protein-DNA-Interaktion`

2.1.3 Immunoprezipitation (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 Prezipitats: 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 Immunoprezipitation

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

polyclonal

Aufbrechen der Proteine in kurze Aminosäureketten (Peptide)

⇓

Peptide in Ratte/Maus geimpft

⇓

extrahieren der B-Lymphozyten aus Serum

⇓

Extrahieren der Antikörper aus den Lymphozyten

⇓

⇓

⇓

ernten verschiedenen Antikörper

⇓

billig

monoclonal

Aufbrechen der Proteine in Peptide

⇓

Peptide in Ratte/Maus geimpft

⇓

extrahieren der B-Lymphozyten aus Milz

⇓

Fusionierung der B-Lymphozyten mit Plasmazellen aus Myelom (Krebszelle - 'unsterblich')

⇓

Hybrid erzeugt (unsterblich + Antikörper)

⇓

Testen der Hybride auf Antigene

⇓

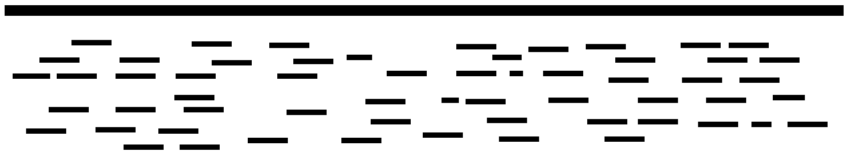
ernten identischer Antikörper

⇓

teuer

3 Peak Calling

Genom:

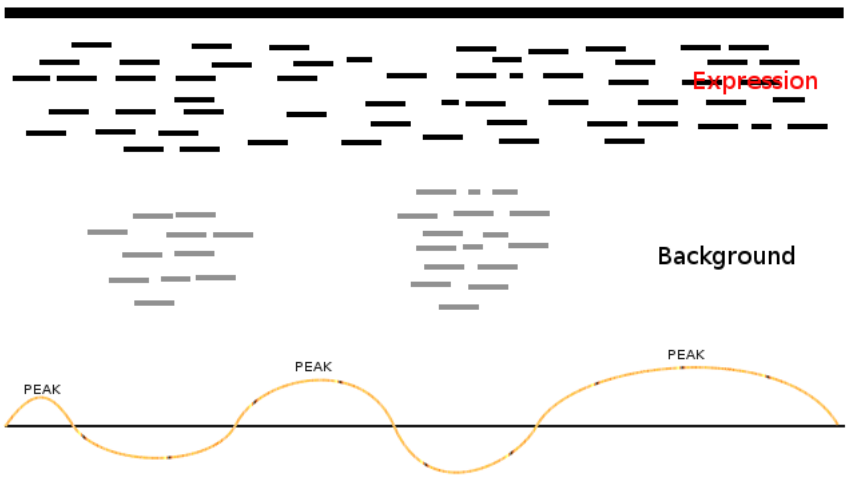


Ergebnis: Sequenziertes Genom/RNA/DNA aus dem Experiment = viele, kurze Reads

Frage: Wo sind die Proteine gebunden?

3 Ansätze:

1. naiver Ansatz: Jedes Nukleotid, dass durch mind. Read bedeckt ist, war gebunden ⇒ viele False Positives, da kurze Reads mehrere Treffer haben können
2. Cut off x: mind. x Reads müssen auf das Nukleotid gemappt sein ⇒ Problem durch sequence bias: Manche Basen einfach zu binden = viele FP
3. enrichment: $log \frac{Expression}{Background}$

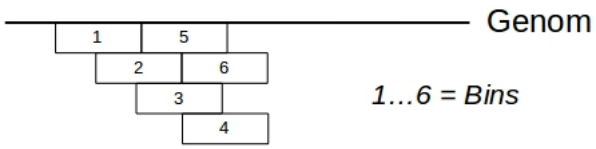


naiv: Wenn Enrichtment > Cutoff → Peak!

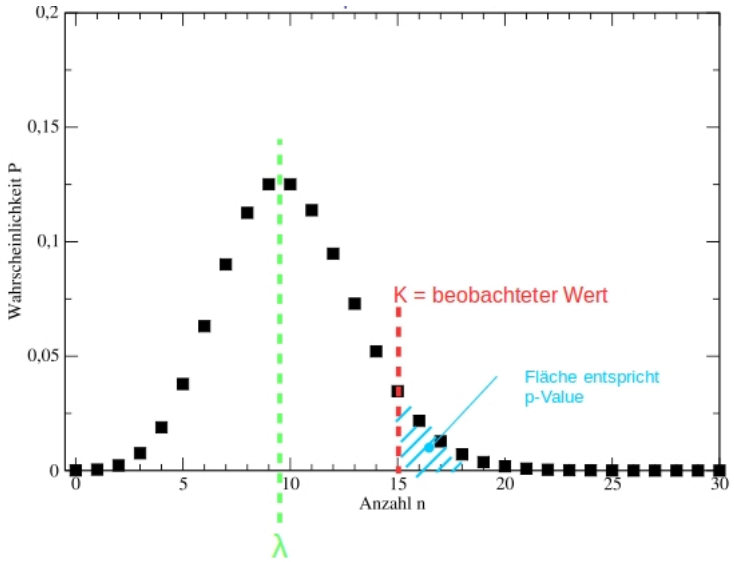
⇒ **daher Entwicklung Peak Calling**[2] - häufig verwendete Software: MACS (Model-based Analysis of ChIP-Seq)[3]

[2] https://en.wikipedia.org/wiki/Peak_calling
[3] http://liulab.dfci.harvard.edu/MACS/

1.
Einteilen des Genoms in Bins (Eimer), n Bins werden Reads eingeordnet
Window: 200 BP und Offset von 1/4 der window size



2.
Zählen der hypothetischen Fragmente pro Bin, +/- Strang
Ergebnis: Liste von Zahlen (Poisson verteilt!)



λ (reelwertig) = Mittelwert der readcounts aus der Hintergrundmessung (Signal aus Experiment nicht zufällig verteilt, daher wird Hintergrund genutzt)
k= Anzahl der reads/fragments aus Experiment

$$P(x \geq k|\lambda) = \sum_{i=k}^{\infty} P\lambda(i) = \underbrace{1 - \sum_{i=0}^{k-1} P\lambda(i)}_{*}$$

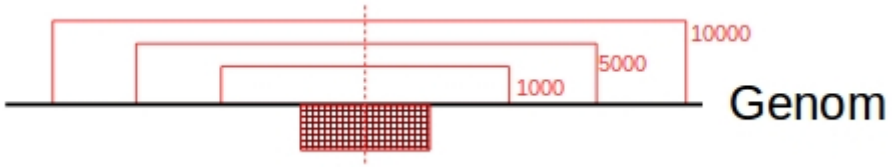
es wird summiert statt integriert, da diskrete Verteilung
* Berechnung der Gegenwahrscheinlichkeit, da $\sum_{i=k}^{\infty}$ schwer zu berechnen

$$P(x \geq k|\lambda) = 1 - \sum_{n=0}^{k-1} \frac{\lambda^n}{n!} e^{-\lambda} \tag{1}$$

P ist global, benötigt wird aber lokal
pro Bin werden 3 lokale + das globale verwendet (Vermeidung von lokalen Sequenzeffekten)

λ_{global} entspricht globalen Mittelwert

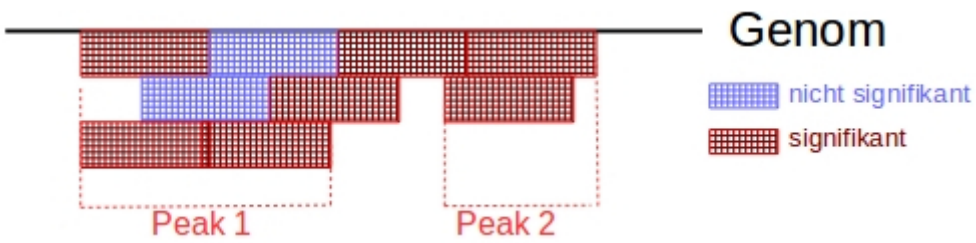
$\lambda_{1k}, \lambda_{5k}, \lambda_{10k}$ $\Rightarrow$ Mittel aller Bins in einem 1k, 5k oder 10k Window zentriert am entsprechenden Bin (K=1000)



λ=max(λglobal, λ1000, λ5000, λ10000)
neues λ verschiebt Mittelwert der Verteilung nach rechts $\rightarrow$ k bleibt gleich $\rightarrow$ Wahrscheinlichkeit sinkt (keine Bias-Unterschätzung) $\rightarrow$ Reduzierung der False Positives

da viele p-Values $\rightarrow$ multiple test problem

3. p-Value correction
- bei MACS: Bonferroni-Holm (bleibt p-Value)
- andere Möglichkeit: q-Value (Storeq) $\rightarrow$ kontrolliert False Discovery Rate
Ergebnis: Signifikanz pro Bin



4. Peakmerging

1. bilden von Peaks über Bereiche von vielen signifikanten Bins
2. Peak-Lücken vermeiden $\Rightarrow$ post processing, wenn Abstand zwischen Peaks $<$ Cutoff $\rightarrow$ Merge Peaks (bei MACS: cut-off=2·Bin-size)

nächster Schritt: Vorhersage durch Motivs (Meme)

- Wo sind die Bindungsstellen?

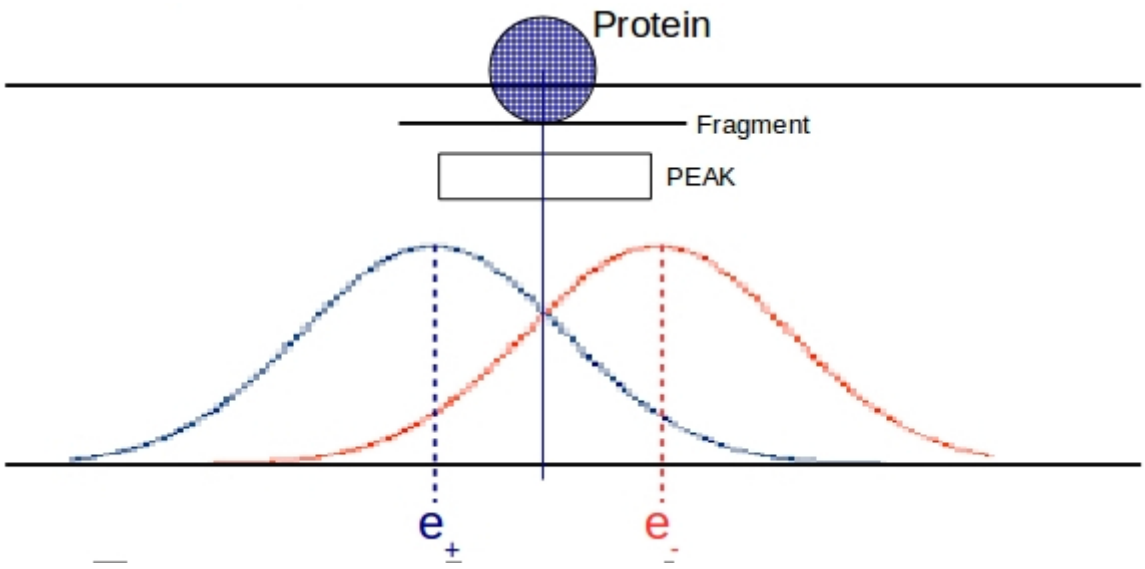
CHIP-Seq:
- Region mit denen das Protein assoziiert ist (nicht wo es gebunden ist!)

$\Rightarrow$ beide Informationen werden benötigt - es werden immer Antikörper benötigt

1. **Protein bekannt**, **gesucht ist RNA** welche vom Protein gebunden wird
   RIP: RNA immunoprecipitation protocol (RIP-seq), Antikörper gegen Protein

2. an welchen Stellen im Genom ist eine **RNA an eine DNA** gebunden?
   Chromatin extrahieren (ChIRP - seq: chromatin isolation by RNA purification), Antikörper gegen RNA oder komplementäres Lesen, DNA sequenzieren

Problem: nur Bereiche (Peaks) der Bindungen ermittelt, keine genauen Positionen



Erwartung: Peak sehr nah um das Protein
Mittelpunkt: e_+ und e_- als Bindungstelle genommen

4 CLIP

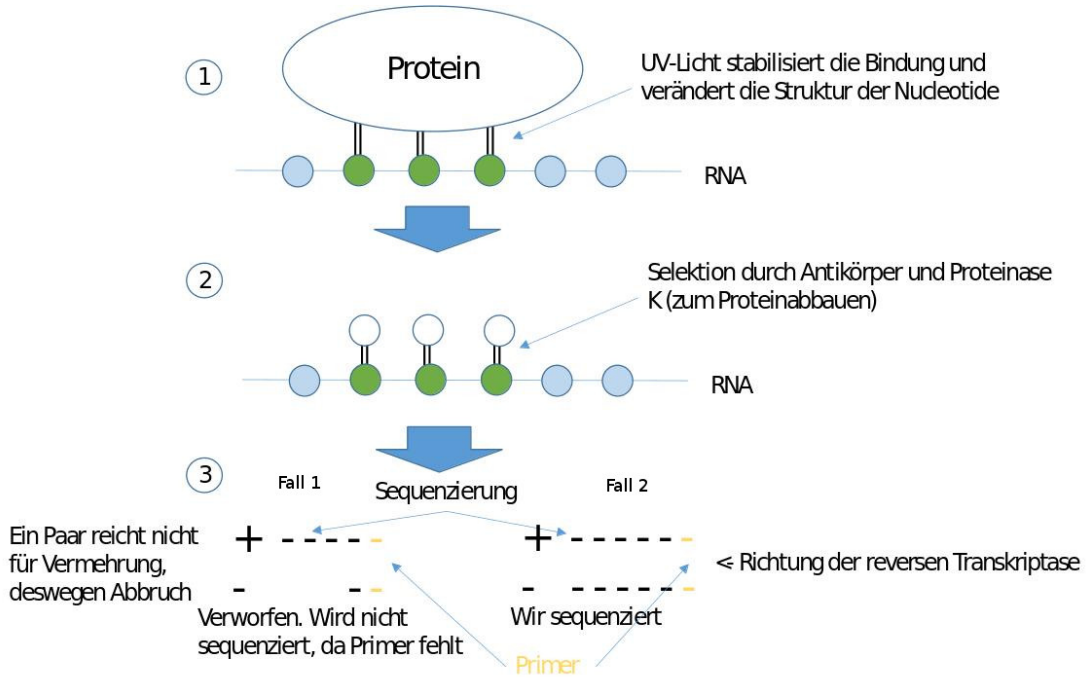
[4]

- Protein an RNA

4.1 CLIP-Seq

cross-linking & immunoprecipitation protocol (cross-linking immunoprecipitation-high-throughput sequencing)[5]

1. Ultravioletes Licht für Cross-Linking, UV-Licht cross linked **NUR** RNA mit Proteinen
2. Induziert UV Mutation der RNA
3. CIMS: Cross-linking induced mutation site

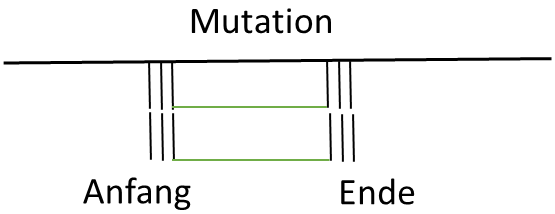


Template ist per Definition Plus-Strang und dass durch die PCR erzeugte der Minus-Strang (auch wenn es in der Zelle anders ist).

- Fall Links und Rechts treten gleichzeitig auf
- wird in vitro gemacht

[4] https://en.wikipedia.org/wiki/CLIP
[5] https://de.wikipedia.org/wiki/CLIP-Seq

Mutation ist Bindungsstelle:



- nach Häufungen schauen
- hohe Sequenziertiefe benötigt

4.2 ICLIP

individual nucleotide–resolution cross-linking and immunoprecipitation protocol[6]

- Schritt 1 und 2 wie bei CLIP-Seq
- Im Schritt 3 werden jetzt aber zirkuläre RNA erzeugt

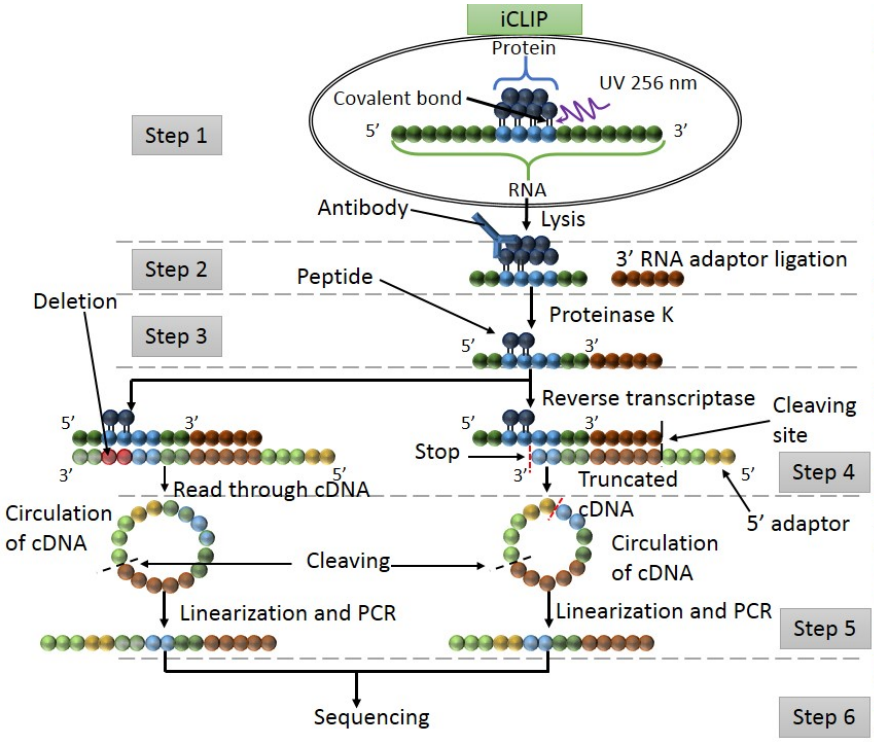


Figure 3: iCLIP
Step 1
Irradiation of the cells with UV light catalyses covalent bond formation between proteins and RNA in direct contact. The cell is lysed, and the protein of interest is isolated using immunoprecipitation.
Step 2
Washing is performed to remove free RNA, and RNA adaptors are ligated at the 3′ ends.
Step 3
Proteinase K digestion is performed. This leaves a peptide at the cross-link site that modifies the chemical structure of the nucleotide.
Step 4
Reverse transcription PCR is performed. This results in both truncated cDNAs and cDNAs that are read through the cross-link sites.
Step 5
Adapters are added to the 5' cDNA ends via circularization. Restriction enzyme cleavage is performed to linearize the cDNAs, allowing both the truncated and read through cDNAs to be sequenced. The position of cDNA truncation allows RNA-Protein interaction sites to be determined at high resolution.
Step 6
Sequencing.

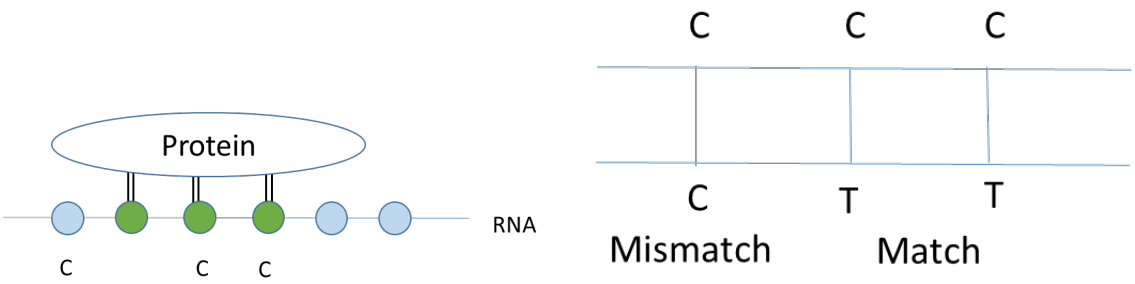
- **Vorteil:** Beide Ergebnisse aus Schritt 5 können in Schritt 6 genutzt werden

[6] https://de.wikipedia.org/wiki/ICLIP

4.3 PAR-CLIP

Photoactivable Ribonucleoside-enhanced CLIP[7]

- Photoaktive (reagieren auf UV-Licht) Ribonucleoside → Diese lassen wir in RNA einbauen
- UV-Licht für Cross Linking
- Cytosin ist photoaktiv, wird durch UV-Licht zu Uracil (welches in DNA zu Thymin wird)
- Antikörper um RNA Fragmente auszuwählen



Seeding: Transkription C→T, Reads C→T

Alignment: angepasste Kostenfunktion(C→T wird nicht bestraft). Dadurch nicht symetrische Kostenfunktion

Vorteil: unterscheiden von verschiedenen Mutationen möglich

[7] https://de.wikipedia.org/wiki/PAR-CLIP

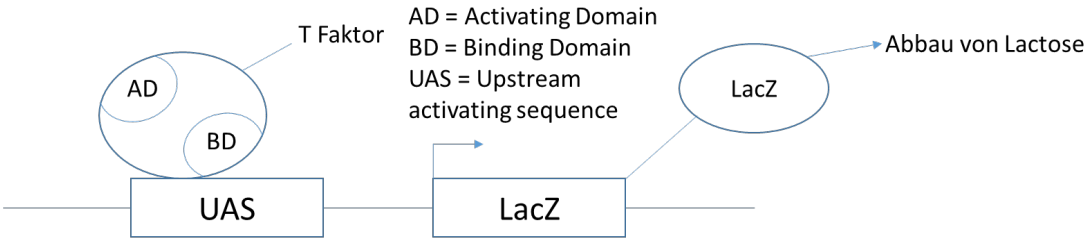
5 Protein-Protein-Interaktion

[8]

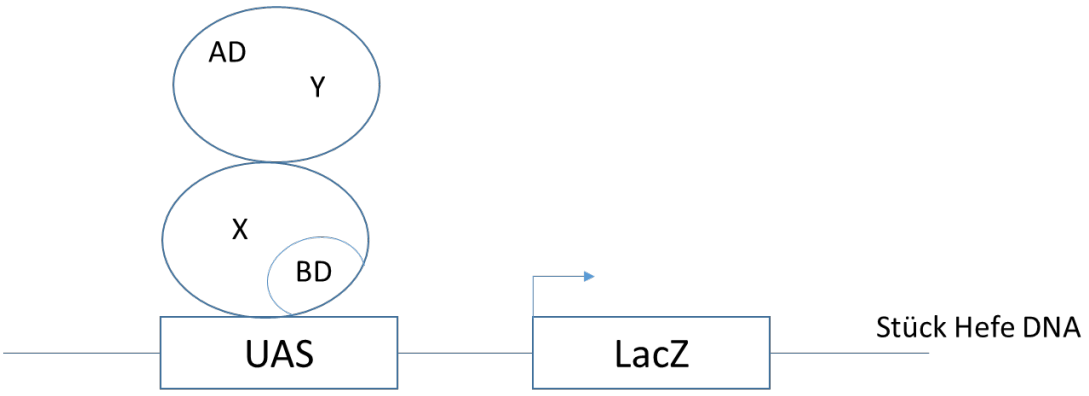
Voraussetzung: beide Proteine bekannt, Wirkung aufeinander soll überprüft/festgestellt werden

Yeast Two-Hybrid System[9]:
- benötigt: Wirtsorganismus: Hefe
- erzeugt: Zwei Hybridproteine

Fragestellung: Interagiert Protein X (Coding DNA X) mit Protein Y (Coding DNA Y)?
Normalfall:



- Coding DNA X und Binding Domain von GAL4 werden in Vektor (=Stück zirkuläre DNA in die ich Proteine binden kann) kloniert
- Vektor mit coding DNA Y & Activating Domain
- Beide Vektoren (haben auch Promotor) werden in Hefezelle eingeschleust



Welche Interaktionspartner hat Protein X?

- Erstellen einer Library (Vektor mit X und Vektor mit einem anderen Protein)
- Da wo Hefe auf Lactose wächst reagieren die Proteine

[8] https://de.wikipedia.org/wiki/Protein-Protein-Interaktion
[9] https://de.wikipedia.org/wiki/Hefe-Zwei-Hybrid-System

Problem:

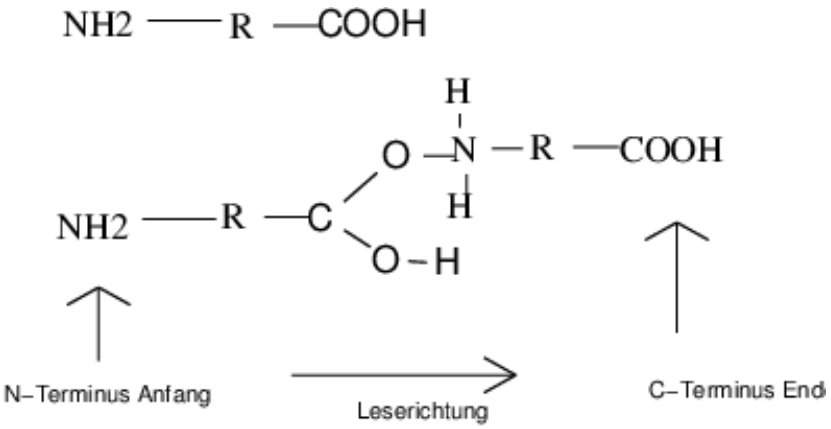
- X&Y falten nicht in natürlicher Struktur und dies führt dazu dass sie nicht mehr binden können (false negativ)
- Das BD oder AD nicht gefaltet werden wie für ihre Funktion notwendig (false negativ)
- Vektor Klonierung funktioniert nicht immer (false negativ)
- Zufällige Aktvierung von Lac Z (false positiv von 80%) $\rightarrow$ man bekommt nur die Info ob Hefe wächst, aber nicht wie stark

In Hefe kennt man ca-50% aller Interaktionen (Interaktom). Mensch: 30% des Interaktom bekannt.

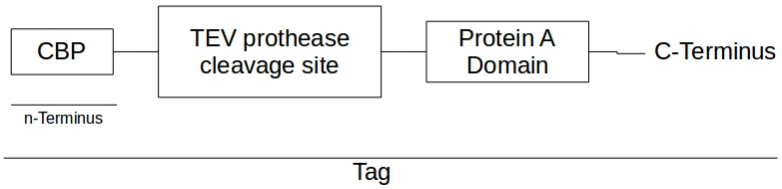
6 Tandem Affinity Purification (TAP)

[10]

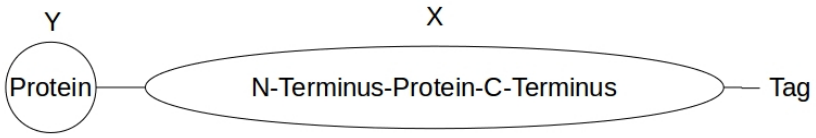
Verwendung: Welche Proteine interagieren untereinander? Suche nach einzelnen Proteinen und Proteinkomplexen



TAP-Tag
- C-terminal variante (es gibt auch n-terminal)



CBP - Calmodulin binding peptide[11]
TEV - tabacco etch virus[12]
IgG - unspezifischer Antikörper (Immunglobulin G)[13]

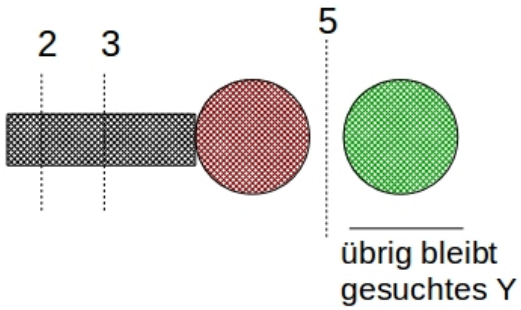


Protein Y wird gesucht! (Tag wird in Plasmid eingeschleußt)

1. Plasmid mit getaggtem Protein & Interaktionspartner werden in Hefezellen inkubiert

[10] https://en.wikipedia.org/wiki/Tandem_affinity_purification
[11] https://en.wikipedia.org/wiki/Calmodulin
[12] https://en.wikipedia.org/wiki/Tobacco_etch_virus
[13] https://de.wikipedia.org/wiki/Immunglobulin_G

2. Affinity purification (Ähnlichkeit Aufreinigung): IgG Matrix bindet die Protein A Domain des Tags (am Ende bleint Tag übrig)
3. mit Hilfe der TEV protease um an TEV protease cleaveage site zu schneiden
4. Calmodium beads um Protein zu extrahieren
5. Auftrennen der Proteine, z.B. durch Ultraschall (nur Interaktionsbindung, keine Peptidbindung!)
6. Indentivizieren von Y durch Massenspektrometer



Probleme:

- durch (häufige) Reinigung hohe Fehlerrate
- durch Tag an Protein Faltung möglicherweise nicht mehr (wie ursprünglich) möglich

weitere Informationsquellen (indirekt)

Ziel: Reduzierung des False-Negatives

- Interaktion über Protein-Protein-Bindungsdomain: Vorhersage über Markovmodelle möglich (Domain, Interaktionspartner)
- Homologie: Vorhersage über Interaktionen in nahen Verwandten
- Textmining auf Publikationen

Filterung

Ziel: Reduzierung der False-Positives

- Co-Expression: werden 2 Proteine gleichzeitig expremiert?
- Lokalisationsinformationen: wenn nicht im gleichen Kompartiment vorhanden, Interaktion nicht möglich

⇒ Ergebnisse durch vorherige Vorgänge: Protein-Protein-Interaktionsnetzwerke in einer Spezies
⇒ **Analyse des Netzwerks**

Protein-Protein-Interaktionsnetzwerk (PPIN) = Graph G = (V,E)
V = Knoten (Proteine)
E = Kanten (Interaktionen) $\subseteq$ VxV $\rightarrow$ erzeugt Paare von Konten
$\rightarrow$ ungerichtete Graphen: (a,b) $\in$ E $\Leftrightarrow$ (b,a) $\in$ E

6.1 Local clique merging algorithm (LCMA)

clique - vollständige subgraphen C
$C = (V', E')$ mit $V' \subseteq V$, $E' \subseteq E$
$\forall x, y \in V' : (x, y) \in E'$

Annahme: dichte Subgraphen repräsentieren Proteinkomplexe

Dichte von G: $\delta(G) = \frac{2 \cdot |E|}{|V| \cdot |V-1|}$

Suche nach dichten Graphen

1. Suchen Knoten u in G mit dem kleinsten Grad (Grad eines Konten = Anzahl der Kanten die von einem Knoten ausgehen)
2. entfernen Knoten (+ Kanten) mit dem geringsten Grad ⇒ erhöht die Dichte in Graphen: $G' = G \backslash \{u\}$
3. wiederhole ab 1 solange gilt: $\delta(G') > \delta(G)$

⇒ lokale Cliquen $C_1, \ldots, C_n$

Merge:
Overlap von $C_x = (V_x, E_x)$ & $C_y = (V_y, E_y)$

$$Overlap = \frac{|V_x \cap V_y|^2}{|V_x| \cdot |V_y|} \tag{2}$$

wenn Overlap > cut-off $C_x \cup C_y = (V_x \cap V_y, E_x \cap E_y)$

Solange wie noch Cliquen gemerged werden & $\underbrace{\sum_n \frac{\delta(n)}{N}}_{averagedensity}$ nicht signifikant schlechter wird (AD' > 0,95 AD)

$\rightarrow$ Verleich mit realen Proteinkomplexen hat gezeigt, das Cliquen keine gute Approximation ergeben

6.2 Clique Finding Algorithzm (CFA)

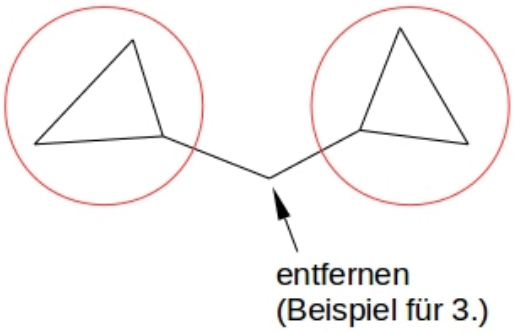
Annahme: Proteinkomplexe k-connected[14]
graphs ⇒ geringe Dichte möglich
k-connected: $k \in \mathbb{N}\ \forall V' \subset V, |V'| < k$, G zusammenhängend
k → Anzahl der Knoten, die entfernt werden können, ohne das G auseinanderfällt
⇒ alle Knoten haben Grad > k

1. entferne alle Knoten mit Grad < k
2. wenn der resultierende Graph weniger als k Knoten hat → kein k-connected Subgraph
3. finden $\{u_1, \ldots, u_n\}$, n<k, so dass $G \backslash \{u_1, \ldots, u_n\}$ nicht mehr zusammenhängen, es entstehen Zusammenhangskomponenten

⇒ für jede Zusammenhangskomponente beginnen bei 1.
wenn $u_1, \ldots, u_n\}$ nicht existiert ⇒ G ist k-connected

[14] `https://de.wikipedia.org/wiki/K-Zusammenhang`

Beispiel: k=2



k $\Rightarrow$ Suche Anzahl n $<$ k = Anzahl der Knoten die entfernt werden können ohne dass der Graph auseinanderfällt

- 1-connected
- 2-connected
- ...
- n-connected

Filtern: dia(G) = Durchmesser von G (Länge des längsten Pfades)
k=1 $\Rightarrow$ dia(G) = 4
k=2 $\Rightarrow$ dia(G) $>$ 2·k
rausgefiltert werden alle dia(G) $<$ 2·k, da dort die Dichte hoch

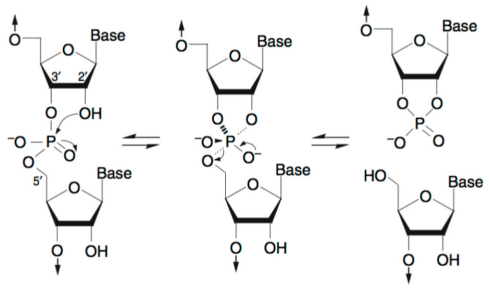
7 RNA structure probing

Bestimmung von:

- Basenpaarung
- Sekundärstruktur und Tertiärstruktur

7.1 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

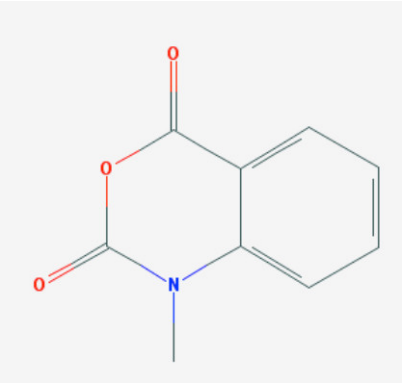
7.2 Chemisches Probing

RNA-modifizierende Chemikalien sind **struktursensitiv** und **sequenzunabhängig**, da sie entweder gepaarte oder ungepaarte Basen modifizieren.

7.2.1 SHAPE-Seq

(**S**elective 2'-**h**ydroxylacetylation **a**nalyzed by **p**rimer **e**xtention **seq**uencing) Mechanismus zur Detektion von Modifikationen

- 2'-OH ist reaktiver wenn die zugehörige Base ungebunden ist
- N-methylisatoic anhybdride (NMIA)reagiert mit 2'-OH-Ende der RNA unter Abgabe von Kohlenstoffdioxid (CO_2).



15

- reverse Transkription: Die RNA wird mit DNA-Molekülen trankribiert. Die Transkription bricht an der NMIA-Bindestelle ab.
- die gewonnenen DNA-Fragmente werden sequenziert und als $Library^+$ gespeichert
- Da es auch zu zufälligen Abbruch bei der reversen Transkription kommen kann, wird als Kontrolle eine $Library^-$ aus einem NMIA-freien Ansatz erzeugt
- Alignment der Reads an Transkriptom der RNA (X_{ij}, wobei i = Basenposition, j = Library)
- Maximum-Likelihood-Model:
  - Positionsraten: $r_i = \frac{x_{i+}}{x_{i-}}$
  - Abbruchrate: Θ
  - simulierte Daten: m_i
- Berechnung der positionsweisen Shape-Reaktivität γ_i

[15] `https://pubchem.ncbi.nlm.nih.gov/compound/N-Methylisatoic_anhydride`

→ Ermittlung der pseudo-Free-Energy

$$\Delta G_{Shape_i} = m * ln(\gamma_i + 1.0) + b \tag{3}$$

m ... Anstieg des Bestrafungswertes
1,0 ... Pseudocount (damit γ auch = 0 sein darf)
b ... negativer Bonus der freien Energie für gepaarte Basen

- Bestimmung der gewichteten freien Energie je Paar:

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

E_{ij} ... Standard Energiemodell

- Berechnung des mfe (??):

$$M_{ij} = min \begin{cases} M(i+1, j \\ \\ min(M(i+1, k-1) * M(k+1, j) * e^{-\frac{E'_{ij}}{kT}}) \end{cases} \tag{5}$$

7.2.2 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($\vec{E}$)

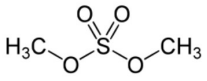
$$\vec{E} = \sum_{\mu} \frac{\varepsilon_\mu^2}{\tau^2} + \sum_{i=0}^{n} \frac{1}{\sigma^2}(p_i(\vec{\varepsilon}) - q_i)^2 \tag{6}$$

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

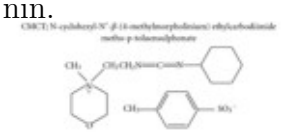
7.2.3 Hydroxyl-Radikal Probing

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 (permanente Herstellung)

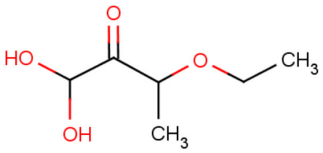
- **DMS**
  Di-Methylsulfat bindet an CH_3 von ungebundenen A bzw. C oder an eines der beiden, wenn sie das letzte Basenpaar einer Helix bilden oder wenn sie direkt neben einem GU-Basenpaar liegen.



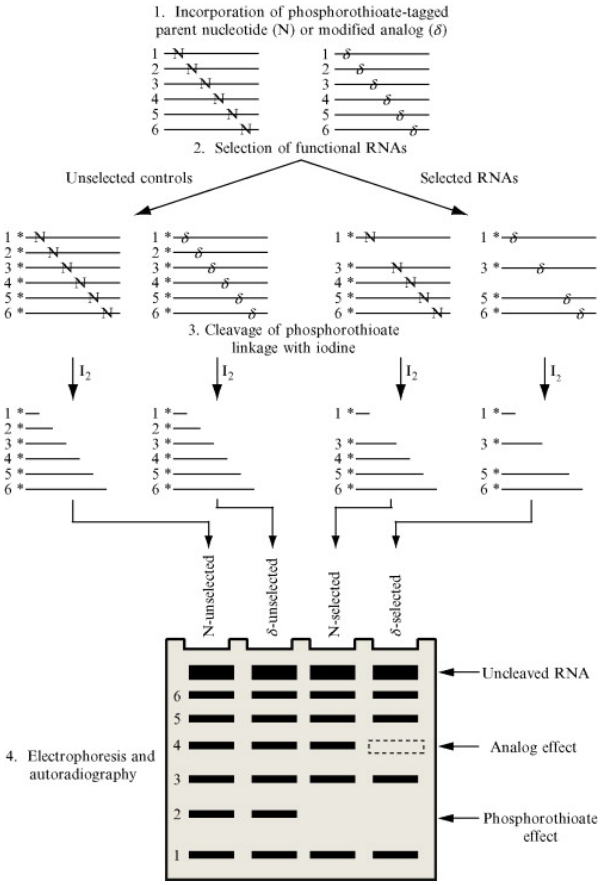
- **CMCT** (1-Cyclohexyl-(2-Morpholinoethyl)Carbodiimid Metho-p-Toluensulfonat) modifiziert vorwiegend ungepaartes Uridin und teilweise ungepaartes Guanin.



- **Kethoxal**
  Kethoxal modifiziert ungepaartes Guanin



7.3 Nucleotide analoge 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 getaggede Analogika[16] und getaggede normale Nukleotide während Transkription markiert
- Annahme: Jedes Transkript hat nur ein getaggedes Nukleotid/Analogon
- Auswahl der aktiven funktionalen RNAs und Erzeugung einer inaktiven Kontrollgruppe (z.B. durch Bindungstest mit Proteinen)
- Cleavage (Beschneiden) hinter der getaggeden Struktur durch Iod (nur die mit Tag!)
- Gelelektrophoresebild → gibt Aussage darüber welche durch Selektion sichtbar werden und welche durch Nukleotid-Einbau sichtbar sind

[16] `https://de.wikipedia.org/wiki/Analogon_(Chemie)`

8 Proteinstrukturen

Methoden

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

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

Nachteil: sehr ungenau und starkes Hintergrundrauschen

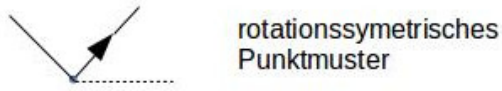
8.1 X-ray crystallography

Voraussetzung: regulären Kristall aus dem Protein

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

X-ray crytallography diffraction:
X-ray → Kristall → Ablenkung
durch Atome → Ablenkung wird durch einen Detektor gemessen
feste Wellenlänge λ, Winkel Θ variieren (Kristall rotieren) → charakteristisches Diffraction pattern → 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 variiren → white x-ray

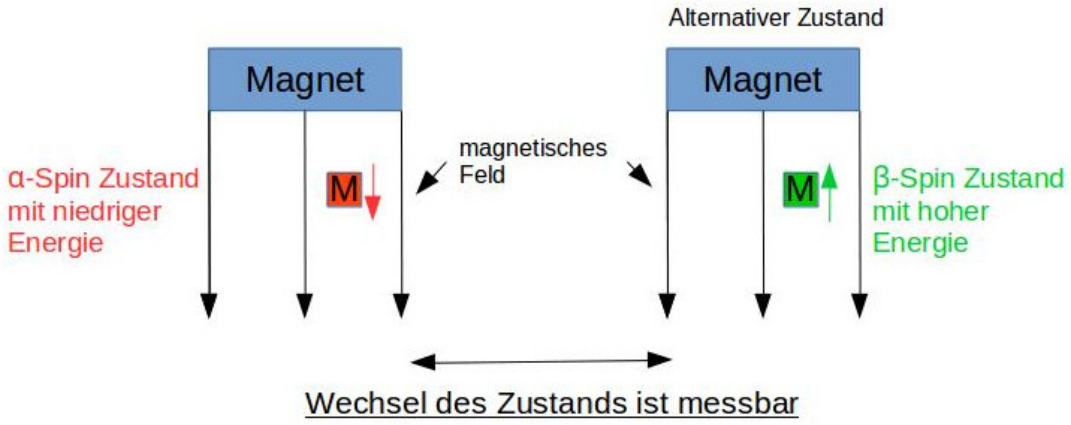


Kombinierte Information aus allen Messungen für verschiedene λ&Θ

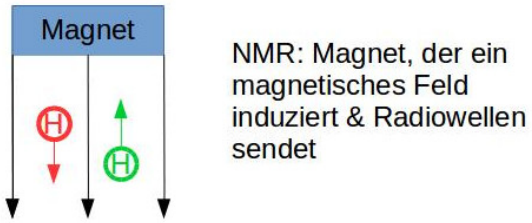
1. Backbone des Proteins ($COOH - NH_2$)
2. Bestimmung der Position der flexiblen Seitenketten der Aminosäuren
3. Verbesserung

8.2 NMR spectroscopy

NMR: nuclear magnetic resonance



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



→ ohne weitere äußere Einflüsse Atom in $\alpha - spin$
→ 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