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 Immunoprezipitation (Selektion mittels Antikörper) . . . 1
2.1.4 Reverse Immunoprezipitation . . . 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) **8**
7.1 Local clique merging algorithm (LCMA) . . . 10
7.2 Clique Finding Algorithzm (CFA) . . . 11

8 RNA structure probing **13**
8.1 objective function approach . . . 13
8.2 Inline-Probing . . . 13
8.3 Chemisches Probing . . . 14
8.3.1 MACS (Model-based Analysis of ChIP-Seq) . . . 14
8.3.2 CLIP: Cross-linking and immunoprecipitation protocol . . . 14
8.3.3 SHAPE-Seq . . . 14
8.3.4 Hydroxyl-Radikal Probing . . . 16
8.3.5 DMS . . . 16
8.3.6 CMCT . . . 16
8.3.7 Kethoxal . . . 16
8.4 Nucleotide analoge interference mapping (NAIM) . . . 17

9 Proteinstrukturen **18**

10 X-ray crystallography **19**

11 NMR spectroscopy **20**

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
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 $\rightarrow$ Länge bestimmt Grad der Zerkleinerung
- 200-1000 BP Fragmente im Idealfall

Ergebnis sind DNA Fragmente mit gebundenen Proteinen

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

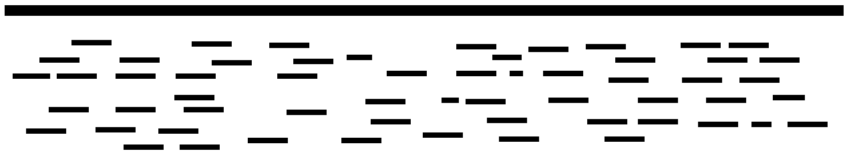
Aufbrechen der Proteine in kurze Aminosäureketten (Peptide)
⇓
Peptide in Ratte/Maus geimpft
⇓
extrahieren der B-Lymphozyten aus Serum
⇓
Extrahieren der Antikörper aus den Lymphozyten
⇓
Antikörper

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 spezifischer Antikörper

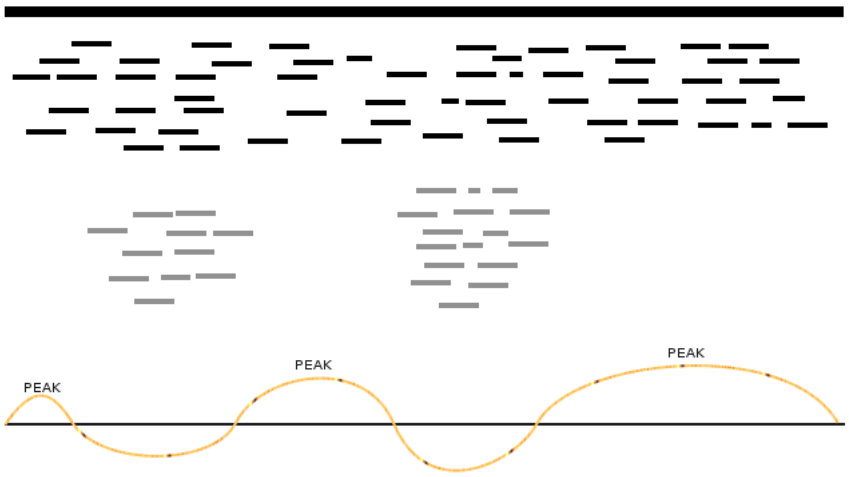
3 Peak Calling

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 Enrichtment > 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:
      λ=max(λglobal, λ1000, λ5000, λ10000)
      → 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 assoziert 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

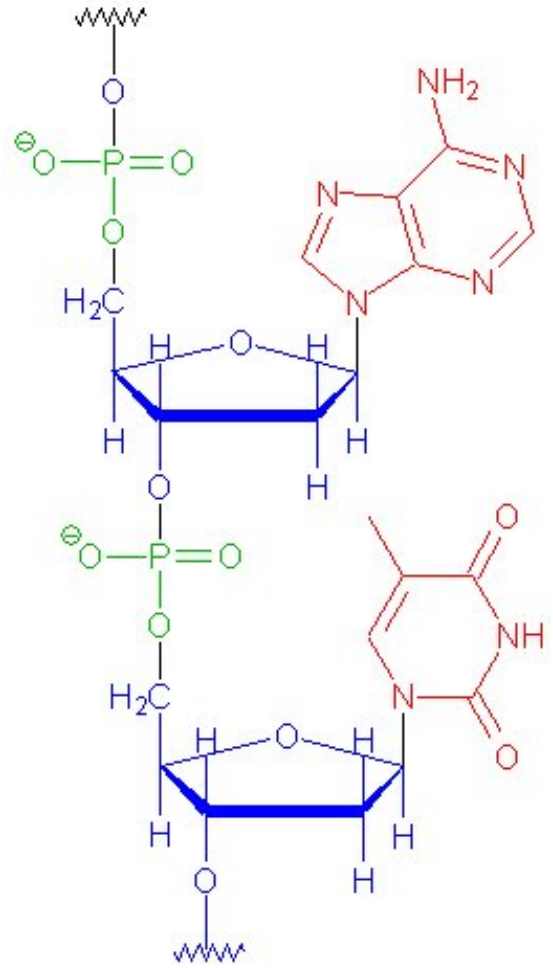
MACS (Model-based Analysis of ChIP-Seq):
1) Einteilen des Genoms in Bins
Window Size: typisch 200 bp & offset (ungefähr 0,25 windows size = 50 bp)
MACS empfiehlt Bin doppelt so groß wie Fragment

2) Zähle die Anzahl an hypothethischen Fragmenten pro Bin (=window)
Fragemente können in mehr als ein Bin fallen

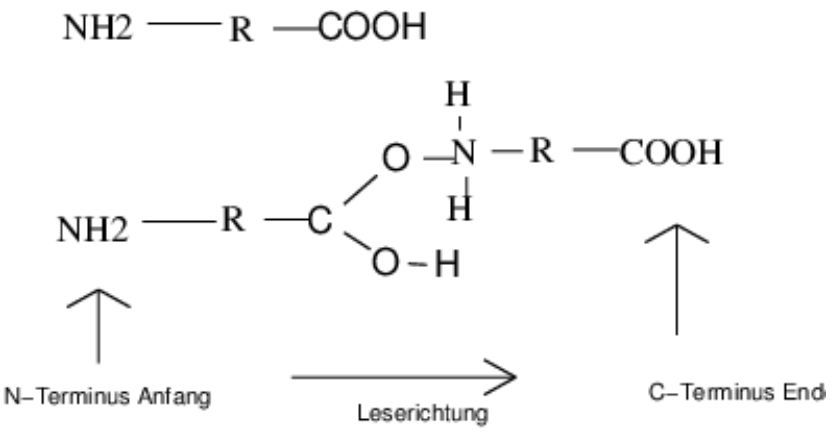
CLIP Cross-linking & immunoprecipitation protocol - Ultraviolettes Licht für cross linking - UV cross linked nur RNA mit Proteinen - induziert UV Mutation der RNA - CIMS: cross-linking induced mutation sites

7 Tandem Affinity Purification (TAP)

[1] **DNA:**

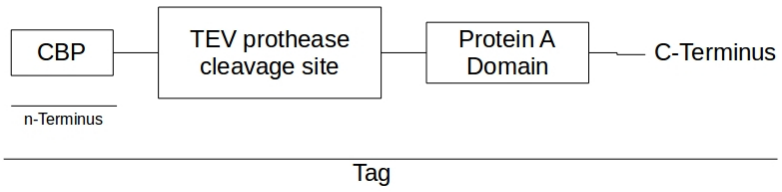


- ähnlich auch bei Proteinen

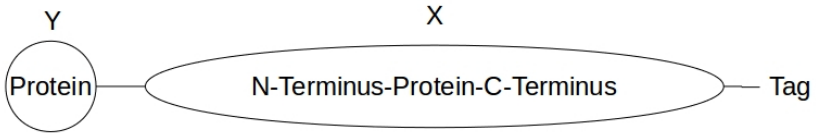


[1] https://en.wikipedia.org/wiki/Tandem_affinity_purification

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

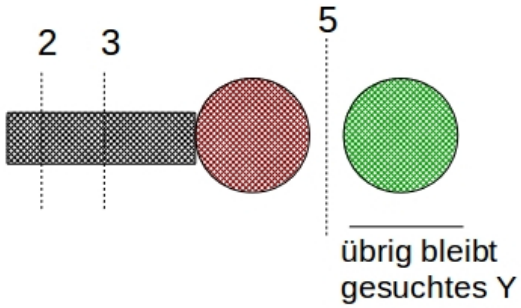


CBP - Calmodulin binding peptide[2]
TEV - tabacco etch virus[3]
IgG - unspezifischer Antikörper[4]



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

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



[2] https://en.wikipedia.org/wiki/Calmodulin
[3] https://en.wikipedia.org/wiki/Tobacco_etch_virus
[4] https://de.wikipedia.org/wiki/Immunglobulin_G

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 → erzeugt Paare von Konten
→ ungerichtete Graphen: (a,b) $\in$ E $\Leftrightarrow$ (b,a) $\in$ E

7.1 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: $\delta(G') > \delta(G)$

$\Rightarrow$ lokale Cliquen $C_1, \ldots, C_n$

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

$$Overlap = \frac{|V_x \cap V_y|^2}{|V_x| \cdot |V_y|} \tag{1}$$

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

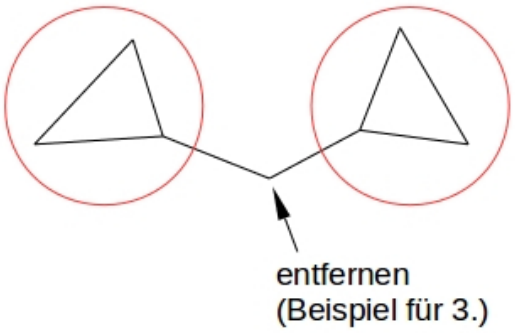
7.2 Clique Finding Algorithzm (CFA)

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

1. entferne alle Knoten mit Grad < k
2. wenn der resultierende Graph weniger als k Knoten hat $\rightarrow$ 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

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

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

8 RNA structure probing

Bestimmung von:

- Basenpaarung
- Sekundärstruktur und Tertiärstruktur

8.1 objective function approach

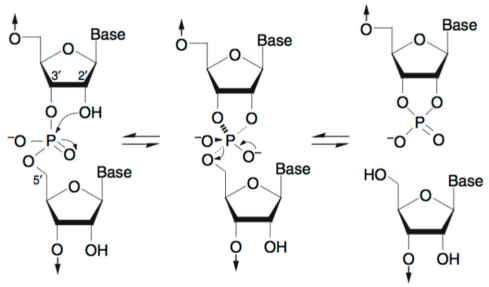
Hard constraints:
→ 3 Aussagen möglich: — = gepaart; . = ungepaart; X = unbekannt
Soft constraints:
→ Wahrscheinlichkeit ob Base an Position Y gepaart ist oder nicht
→ Minimiere den Fehler F($\vec{E}$)

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

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

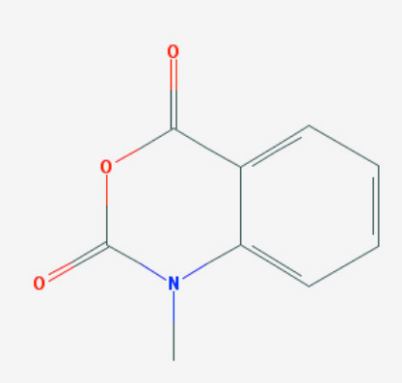
- Ultraviolettes Licht für cross linking
- UV cross linked nur RNA mit Proteinen
- induziert UV Mutation der RNA
- CIMS: cross-linking induced mutation sites

8.3.3 SHAPE-Seq

(**S**elective 2'-**h**ydroxylacetylation **a**nalyzed by **p**rimer **e**xtention **seq**uencing)

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

- 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 trankribiert.I m 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

$$\Delta G_{Shape_i} = m * ln(\gamma_i * 1,0) + b \tag{3}$$

m ... Anstieg des Bestrafungswertes
1,0 ... Pseudocount b ... negativer Bonus der freien Energie für gepaarte Basen

$$M_{ij} = min \begin{cases} M(i+1,j \\ \\ min(M(i+1,k-1) * M(k+1,j) * e^{-\frac{E'_{ij}}{kT}} \end{cases} \tag{4}$$

wobei:

$$E'_{ij} = E_{ij} + \Delta G_{Shape_i} + \Delta G_{Shape_j} \tag{5}$$

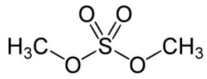
E_{ij} ... Standard Energiemodell

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

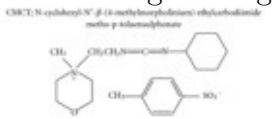
8.3.5 DMS

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.



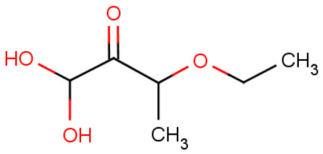
8.3.6 CMCT

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

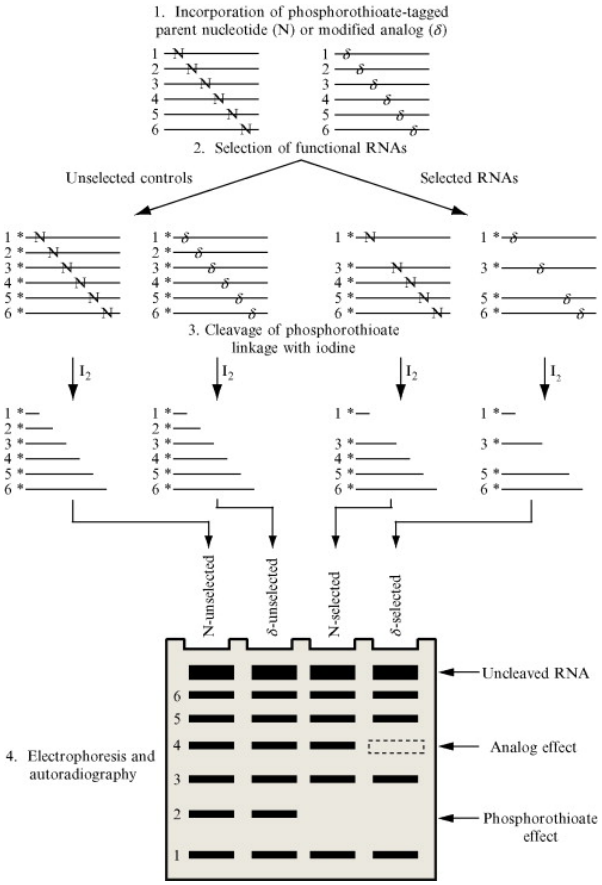


8.3.7 Kethoxal

Kethoxal modifiziert ungepaartes Guanin



8.4 Nucleotide 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 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
- Cleavage (Beschneiden) hinter der getaggeden Struktur durch Iod
- Gelelektrophoresebild → gibt Aussage darüber welche durch Selektion sichtbar werden und welche durch Nukleotid-Einbau sichtbar sind

9 Proteinstrukturen

Methoden

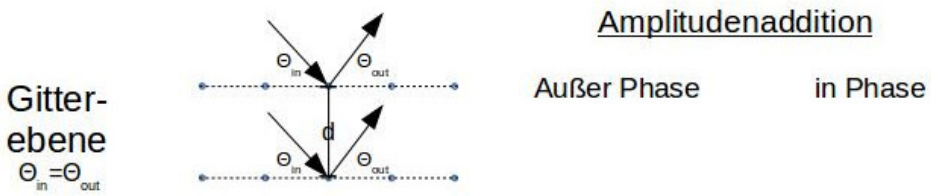
- NMR-Spektroskopie (Protein in Lösung)
- Röntgen-Kristallographie (Protein als Kristall)

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

Nachteil: sehr ungenau und starkes Hintergrundrauschen

10 X-ray crystallography

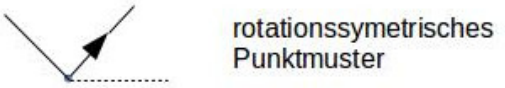
Voraussetzung: regulären Kristall aus dem Protein



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

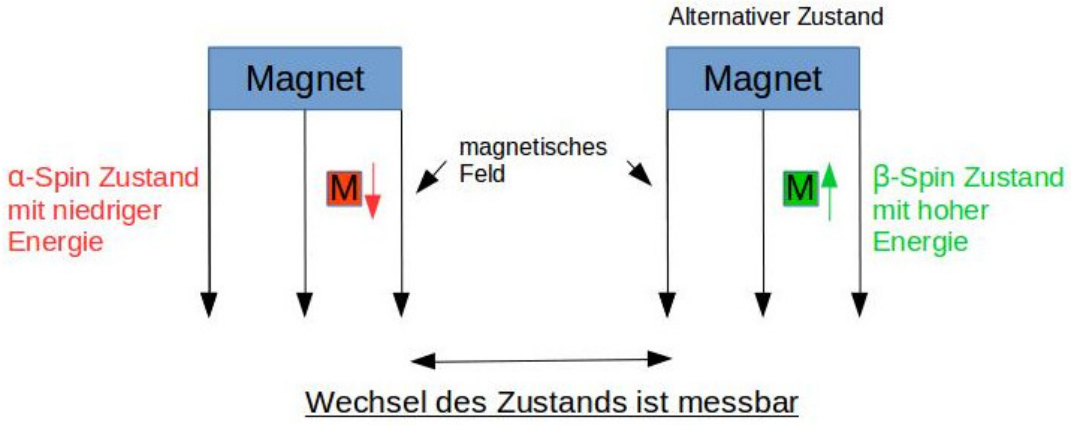


Kombinierte Information aus allen Messungen für verschiedene λ&Θ

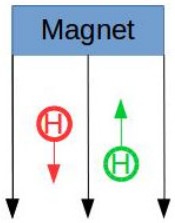
1. Backbone des Proteins ($COOH - NH_2$)
2. Bestimmung der Position der flexiblen Seitenketten der Aminosäuren
3. Verbesserung

11 NMR spectroscopy

NMR: nuclear magnetic resonance



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



NMR: Magnet, der ein magnetisches Feld induziert & Radiowellen sendet

→ 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