

Vorhersage von miRNA targets in Säugetieren

- miRNA target prediction -**
- am Beispiel von TargetScan -**

Christoph Theunert

8. Fachsemester Bioinformatik

christoph.theunert@web.de

Gliederung

- I miRNA Eigenschaften
- II miRNA Funktionen
- III Algorithmus Vorinformationen
- IV → Schritte
- V → Ergebnisse
- VI → Nachteile
- VII Beobachtungen
- VIII Neuerungen /
Verbesserungen

miRNA - Eigenschaften

- miRNAs ~22nt lang, Pflanzen- und Tierzellen
- von Dicer (Proteinkomplex) aus dem hairpin einer endogenen RNA prozessiert
- miRNAs liegen in der Zelle als RNP assoziiert mit Proteinen vor
- wichtige Genregulatorische Eigenschaften
- binden an mRNAs Proteinkodierender

miRNA – Funktionen

- Markierung von Genen für posttranskriptionelle Repression (keine Translation)
- miRNA targets sind sehr häufig in Genen für Transkriptionsfaktoren eingebettet
- Zellproliferation
- Zelltod, Fett Metabolismus in Fliegen
- Blatt - und Blütenentwicklung in Pflanzen
- ca. 0.5 -1 % aller Gene in Menschen, Fliegen und Würmern sind scheinbar miRNA Gene
- →Annahme: wesentlich mehr regulatorische Funktionen, die noch nicht entdeckt sind

miRNA - Funktionen

- Vermutung, dass viele Säugetier miRNAs wichtige Rollen bei der Entwicklung und anderen Prozessen spielen
- → wird unterstützt durch ihre gewebsspezifischen und entwicklungsstufenspezifischen Expressionsmuster
- außerdem durch evolutionär konservierte Bereiche (oftmals in Säugetieren, und Invertebraten Homologen)
- Aufgaben in RNAi specificity complex (RISC)
- für Vertebraten – Funktionen bis dato nur teilweise erforscht
- aber Vermutungen besagen, daß sie einen großen Teil des Transkriptoms regulieren (Gesamtheit aller in einer Zelle transkribierten Gene)

miRNA - Funktionen

- in Pflanzen wesentlich leichter miRNA targets zu identifizieren, sie binden fast perfekt an ihre targets
- 'relativ simples' Suchen nach Komplementarität von miRNA und mRNA (signal noise ratio von z.T. 10:1)
- tierische miRNAs binden weniger stark, z.T. mit einigen ungebundenen Nukleotiden
- somit ergeben sich dabei wesentlich kompliziertere sek. Strukturen
- es konnte gezeigt werden, daß bei Drosophila und c.elegans durch diese Methode nicht mehr targets identifiziert werden als durch puren Zufall gefunden werden

Algorithmus - Vorinformationen

- **Ziel**→
- Vorhersage von Zielgenen für konservierte miRNAs in Wirbeltieren (Vertebraten)
- **Prinzip**→
- Identifizierung der mRNA, die eine konservierte Bindung mit "miRNA seed" eingeht
- kombiniert thermodynamisch basierte Modelle von RNA:RNA Duplexen mit vergleichenden Sequenzanalysen, um miRNA targets vorherzusagen, die über mehrere Genome hinweg konserviert sind
- dabei scheint die Bindung der mRNA an die Nt 2..8 der 5' miRNA von hoher Bedeutung zu sein

Algorithmus - Vorinformationen

- Input→
- 1 miRNA, in mehreren Organismen konserviert
- Menge an orthologen¹ 3'UTR² Sequenzen dieser Organismen
- Homologe³ gefunden mittels BLASTN und MiRscan
- ¹ ortholog=Gene in unterschiedlichen Spezies, die funktionell verwandt sind und von einem gemeinsamen Vorläufer abstammen
- ² UTR =untranslated region (werden transkribiert, aber nicht translatiert, in ihnen sind wichtige Informationen für die OpenReadingFrames enthalten)
- ³ homologe Gene=Gene, die in der Sequenz ihrer Basenpaare eine hohe Übereinstimmung aufweisen und mutmaßlich aus einer Ursequenz (oder Ur-Exons) hervorgegangen sind

Algorithmus - Vorinformationen

orthologe 3' UTR Sequenz der target
mRNA

"seed matches" (WC-
Komplementarität)

Cap ——— ACAAAGCUGGAGACAAUGCGAUCCCUUUGGAUGUCUUCUUG ——— poly (A)

3' CUAGUUACGCUAGGGAAACCU 5'

"miRNA seed" (5' konservierte
Region, nt 2..8)

Algorithmus - Schritte

- |
- Durchsuchen der 3'UTR Sequenzen der Organismen nach Segmenten perfekter Watson Crick Basenkomplementarität zu den Basen 2-8 der 5' miRNA

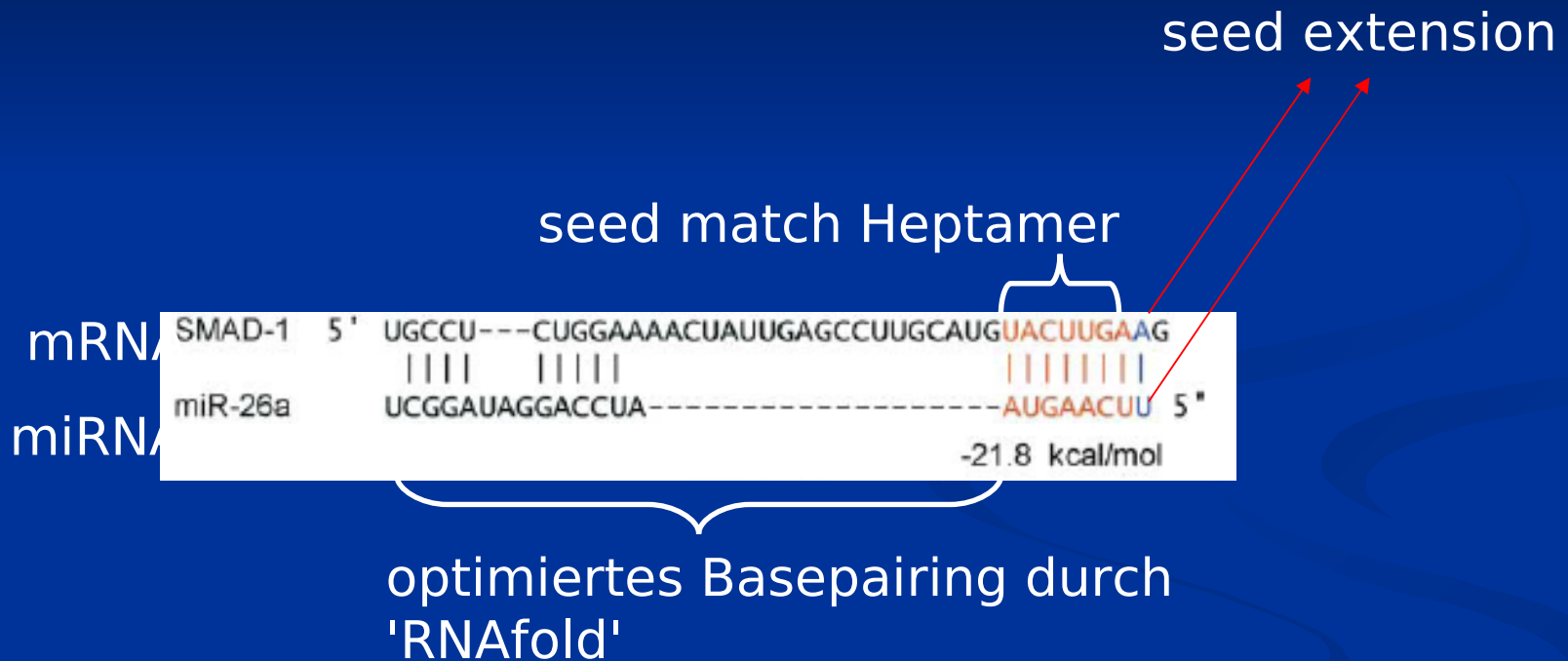
Algorithmus - Schritte

- II
- erweitern jedes seed matches mit zusätzlichen Basenpaaren zur miRNA soweit wie möglich in jede Richtung
- erlaubt sind G:U Paare
- Abbruch bei Mismatches bzw. Gaps
- somit normalisiert man die Sequenzen

Algorithmus - Schritte

- III
- "RNAfold" liefert optimierte Basenpaarung für das 3' Ende der miRNA zu den Basen der UTR, die direkt 5' jedes seed matches lokalisiert ist
- damit verlängert man jeden seed match zu einer längeren target Sequenz

Algorithmus - Schritte



Algorithmus - Schritte

- IV
- jedem miRNA:target Duplex wird eine freie Energie G zugewiesen
- mittels 'RNAeval' (Hofacker, Stadler)
 - RNAeval berechnet die freie Energie eines RNA Moleküls mit gegebener Sekundärstruktur

Algorithmus - Schritte

- V

- jede UTR erhält einen Z Score $Z = \sum_{k=1}^n e^{-\frac{G_k}{T}}$

- n = Anzahl der seed matches pro UTR

- G_k = freie Energie der miRNA:target Interaktion [Kcal/mol]

- k = kte target site, die im vorigen Schritt bewertet wurde

- T = Gewichtungsfaktor, verbindet Z-Score und vorhergesagte freie Energie

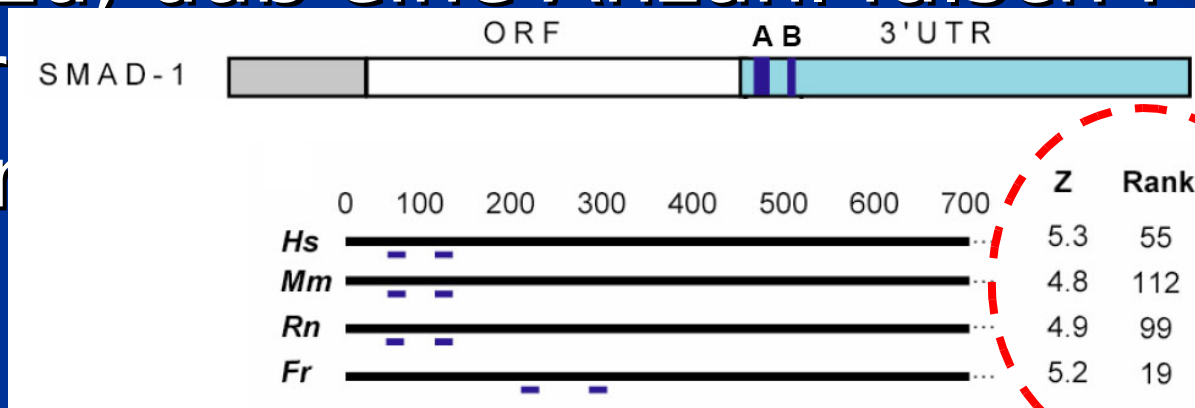
- gewichtet Affinität der Basenpaarung

- Wert beeinflusst die relative Gewichtung der UTRs

- UTR's ohne seed match bekommen Z=1.0

Algorithmus - Schritte

- VI
- sortieren der UTRs eines Organismus nach Z Score und Zuweisung eines Ranges R_i
- Z und R dienen dann unter anderem dazu, daß eine Anzahl falsch Positiver



Algorithmus - Schritte

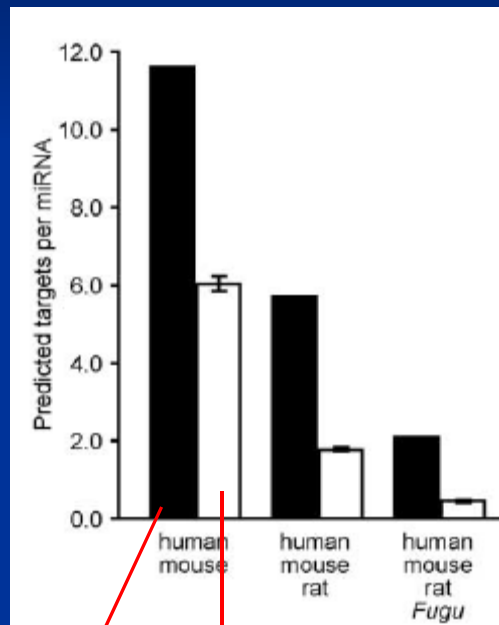
- VII
- wiederholt den beschriebenen Prozess für die Menge an UTRs in jedem Organismus

Algorithmus – Schritte

- VIII
- wählt als targets diejenigen Gene aus, für die
 - $Z_i \geq Z_c$ und $R_i \leq R_c$
- Z_c = vorausgewählter Maximalwert für Z score
- R_c = vorausgewählter Maximalwert für Rang
- R_c , Z_c und T sind frei bestimmbar
- T und Z kommt keine thermodynamische Bedeutung zu sondern sie sind ein Mittel

Algorithmus - Ergebnisse

- Parameter $R_c=200$, $Z_c=4.5$, $T=20$



Signal

Noise

- Anzahl der vorhergesagten targets pro miRNA in verschiedenen Genomen

- Signal sind von targetScan vorhergesagte miRNA targets

- noise sind Anzahl der targets, welche für shuffled miRNAs vorhergesagt wurden (falsch positiv)

- zufällig vertauschte miRNA Sequenzen, die sich nicht mehr als 15% von der ursprünglichen miRNA unterscheiden

- d.h. zur Kontrolle werden signal und noise einander verglichen

- im Endeffekt :

- 3.9 wahre targets pro miRNA in Säugern

- signal:noise ratio von 3.5

- je mehr Genome man in die Berechnungen einbezieht, umso besser wird die Signal : noise ratio

- → Wichtigkeit der evolutionären Konservierung bei diesem Ansatz

- allerdings nimmt dabei auch die Anzahl der vorhergesagten targets pro miRNA ab

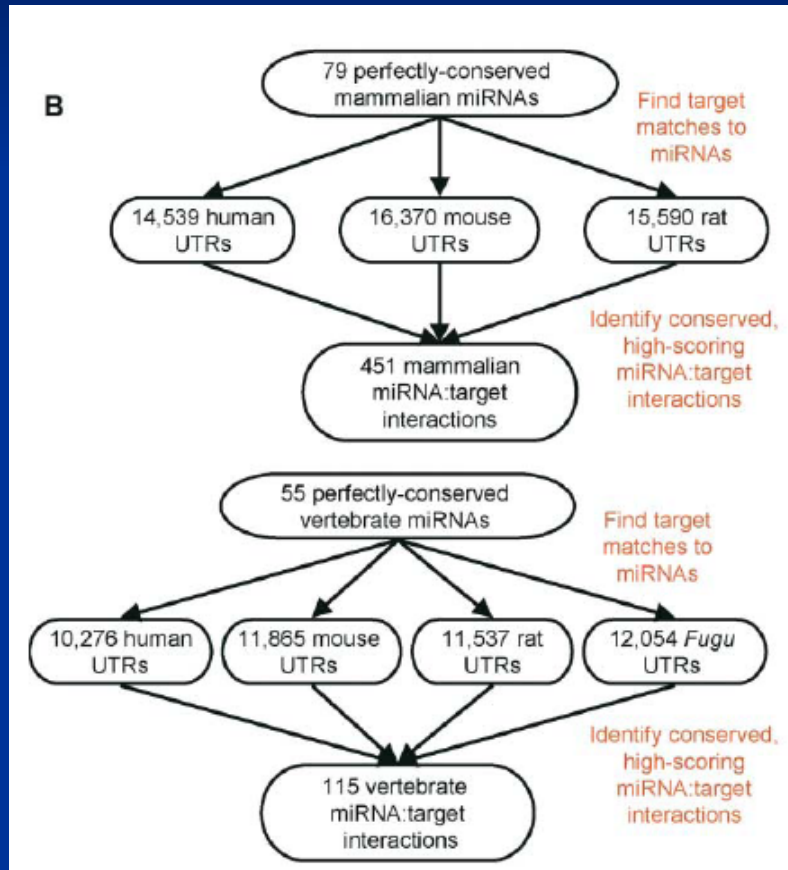
- z.B. kann es sein, daß orthologe Gene in den Annotationen fehlen, je mehr Genome in Betracht gezogen werden

- oder einige miRNA : mRNA Interaktionen sind nicht

Algorithmus - Ergebnisse

- Signal wird verglichen mit Anzahl von targets die für Kohorten von randomisierten miRNAs vorhergesagt wurden (Ermittlung der false positive)
- die mutierten miRNA Kohorten wurden durch 'MiRshuffle' erzeugt
- es entsteht dabei für eine gegebene miRNA eine Reihe von zuf. permutierten miRNAs mit gleicher Länge und Basenkompositionen
- sie wird in Betracht gezogen, wenn bestimmte vorher festgelegte Kriterien erfüllt werden

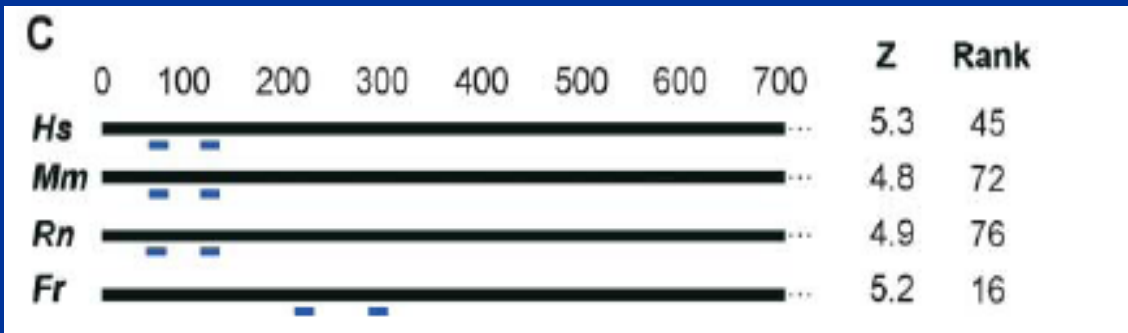
Algorithmus - Ergebnisse



Identifizierung von targets konserviert in Säugern (A) und in Säugern und Fischen (B)

angegeben ist die Anzahl der Gene jedes Organismus' mit identifizierten Orthologen in jedem anderen Organismus

Algorithmus - Ergebnisse



Positionen zweier target sites für die miR-26a (blau) in orthologen SMAD-1 3' UTR Sequenzen von Menschen(Hs), Maus(Mm), Ratte(Rn), und Fugu(Fr, Kugelfisch) mit Z score und Rang jedes miRNA:UTR Paares

Algorithmus - Nachteile

- unvollständige Annotationen für orthologe UTRs
- reale Targets fallen durch Z score und Rang Kriterien
- Bindestellen außerhalb 3' UTR unbeachtet
 - oft bei Pflanzen beobachtet
- nur konservierte Gene in Betracht gezogen
- gleichzeitige Interaktionen verschiedener miRNAs
 - an einer UTR ausgeschlossen
- Vermutung:
- → reale Anzahl an Zielgenen pro miRNA ist

Beobachtungen

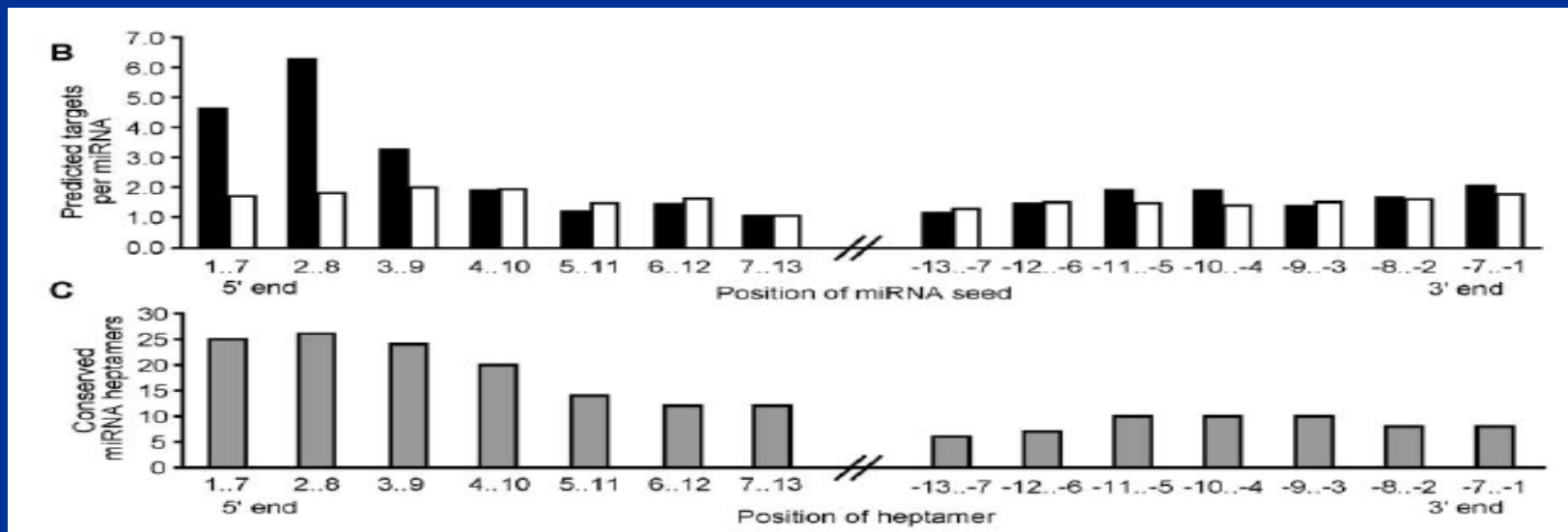
- 5' und 3' Ende der miRNAs werden von targetScan unterschiedlich behandelt
- Forderung: perfekte BP am 5' seed, aber keine solche Forderung am 3' Ende
- die Wichtigkeit der 5' BP wurden in früheren Beobachtungen bestätigt (lin-14 mRNA mit lin-4 miRNA)
- 5' Enden von verwandten miRNAs scheinen besser konserviert zu sein als die 3' Enden
- die höchsten signal:noise ratios wurden beobachtet wenn der seed die nt 2..8 abdeckt
- →stützt die These der Wichtigkeit der 5' BP

Beobachtungen

- die Anzahl an vorhergesagten targets ist am größten bei hoch konservierten miRNAs
- je stärker konserviert eine miRNA ist, umso höher auch das targetScan signal

Beobachtungen

- B – die größte Anzahl an predicted targets ist im Bereich der Nukleotide 2...8 des 5' Endes
- C -die miRNA seeds im Bereich der der Nukleotide 1...7 bzw. 2...8 des 5' Endes sind am stärksten konserviert

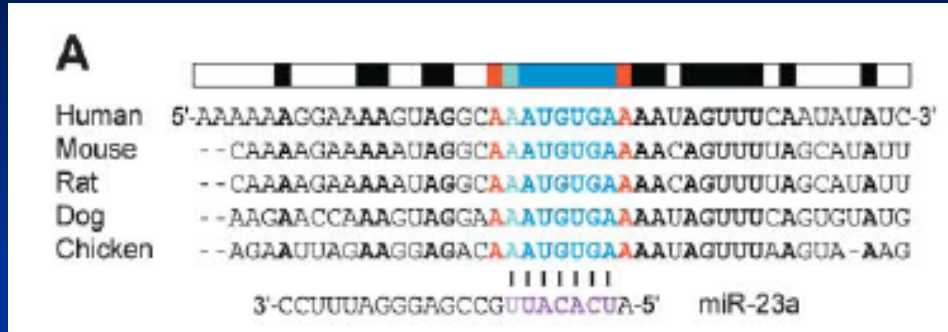


- die Beobachtung der in vitro Paarung an Nt 2..8 scheint auch die Wichtigkeit für die in vivo targetPrediction anzugeben

Neuerungen / Verbesserungen

- mit besser angepassten Genomen kann man die Notwendigkeit der score und rank Schwellwerte abschwächen und zT eliminieren
- außerdem konnte man die Größe des Matches mit der seed Region der miRNA von 7 auf 6 nt reduzieren
- es konnte gezeigt werden, dass mit gut aufbereiteten Daten ein target nur mittels des 6 nt langen seed matches zur miRNA in orthologen UTRs aller Testgenome gefunden werden kann

Neuerungen / Verbesserungen



mRNA

"seed-matches"

miRNA "seed"

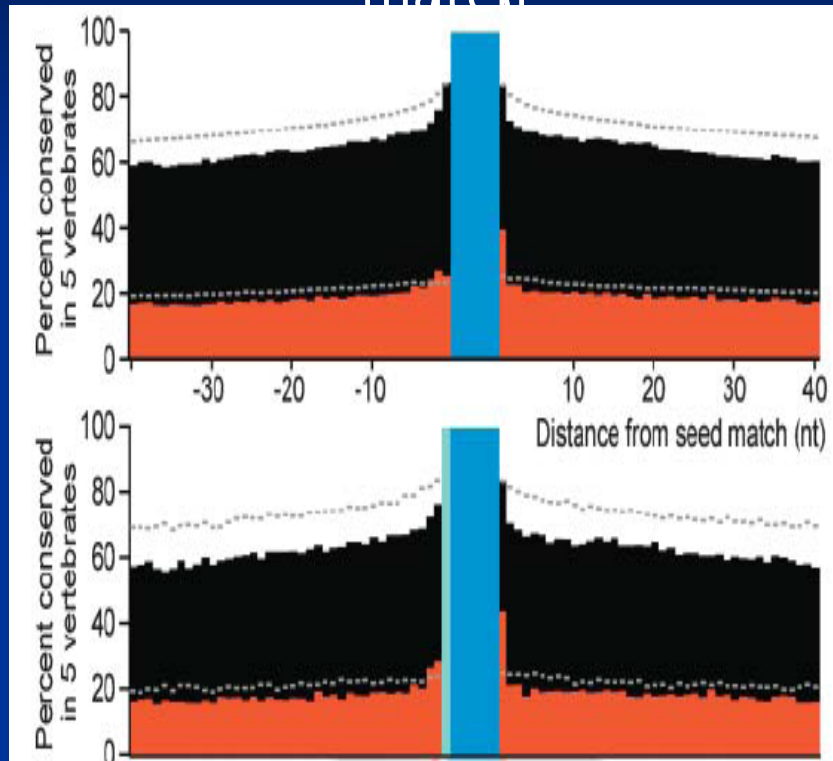
- Alignment von orthologen Segmenten der HIC UTR
- rot markiert sind konservierte, den "seed-match" flankierende Adenosine
- da man heraus fand, daß oftmals der 3' Terminus des seed match ein Adenosin ist, bzw. mehrere A's, stellte man fest, daß eine nur 6 statt 7 lange nt Sequenz, gefolgt von diesem A, die target Spezifität erhöht
- auch die Upstream (zum 5' Ende hin) Regionen des seed matches sind oftmals hoch konserviert
- die Suche nach dieser 6 nt + A langen Sequenz erhöht die signal:noise ratio auf 2,8:1

Neuerungen / Verbesserungen

- es wird spekuliert, daß dieses A am 3' Ende der mRNA vom Silencing Complex erkannt wird, was zu Interaktionen zwischen miRNA und target führt
- fasst man diese neuen Erkenntnisse zusammen, kann man die s:n ratio auf 5.6:1 verbessern
- neue Erkenntnisse werden in "targetScanS" angewendet (Unterschied u.a. 6 nt seed Paarung)
- d.h. targetScanS sagt targets vorher, die entweder ein m8 match oder ein t1 'A'

Neuerungen / Verbesserungen

seed
match



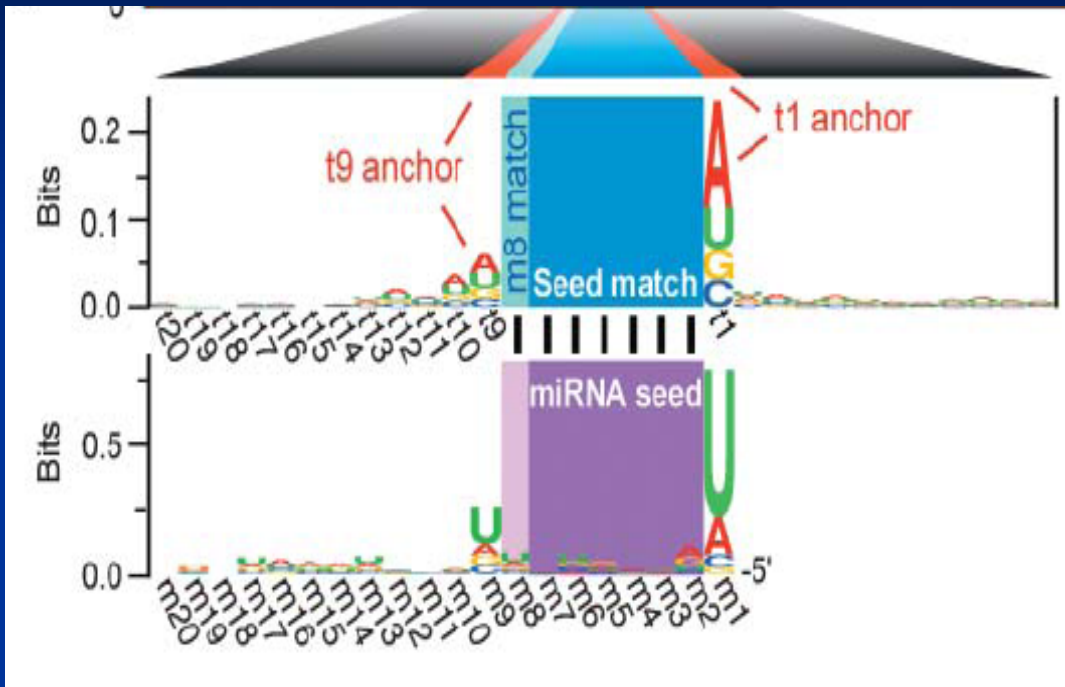
flankierende Regionen

- für jede Position, die den konservierten seed match (blau) flankiert, ist prozentual angegeben, in wie vielen seed matches der untersuchten Spezies diese Positionen auch konserviert sind

- schwarze Balken: Konservierung einer der 4 Möglichkeiten

- roter Balken: Konservierung eines

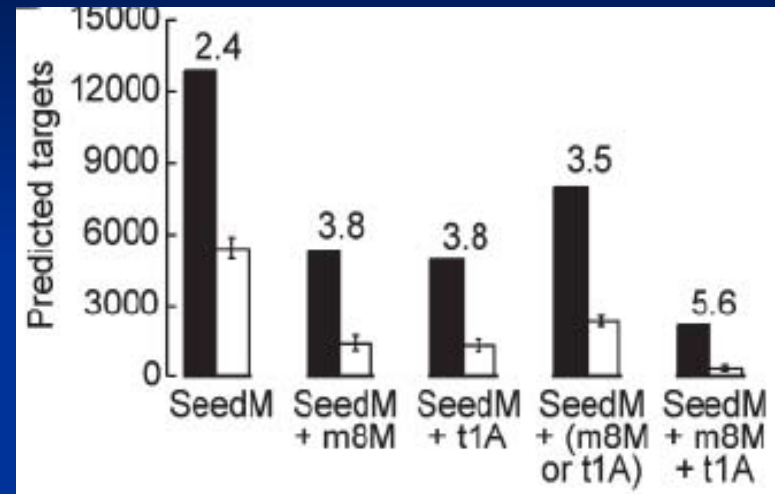
Neuerungen / Verbesserungen



- Sequenzidentität der direkt die seed matches flankierenden Nukleotide
- 6nt langer seed match
- m8 match
- t9, t1 anchor
- gemessen relative Entropie relativ zur Hintergrund

Neuerungen - Verbesserungen

- man kann beobachten, dass miRNA targets manchmal nicht über die 8nt lange center Sequenz des seed match hinaus konserviert sind
- daher stellte man eine weitere Bedingung auf:



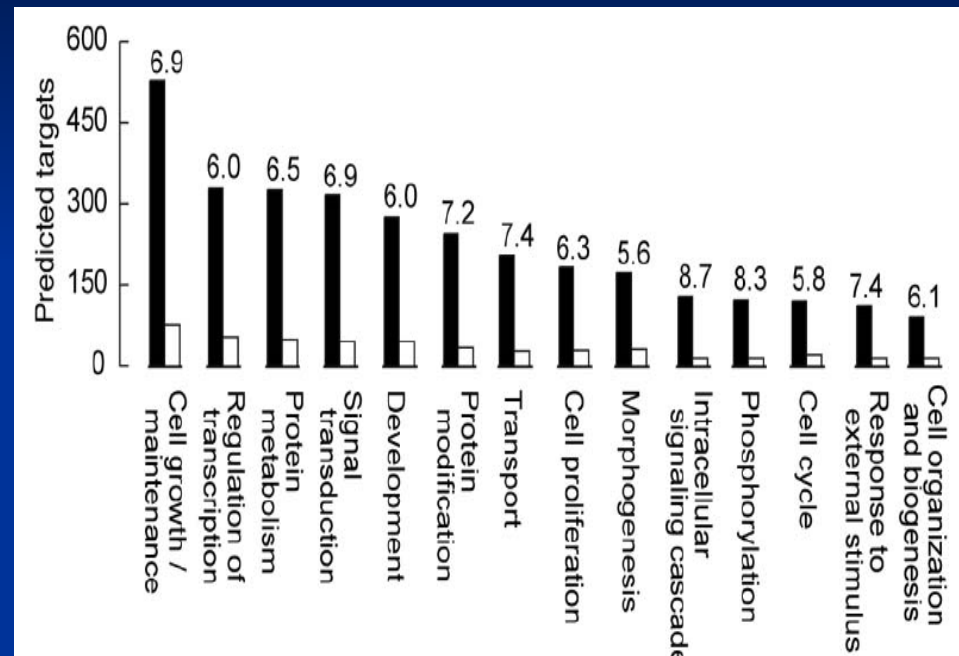
- seed matches müssen in kleine konservierte 'Inseln' fallen → erhöht die signal : noise ratio

Neuerungen / Verbesserungen

- → targetScan / targetScanS beachten keine targets in denen wobbles oder mismatches auftreten
- wird geduldet, da erfahrungsgemäß somit die signal:noise ratio gegen 1:1 gehen würde

Neuerungen / Verbesserungen

- wendet man targetScanS auf 5' UTR Sequenzen an, erhält man nur schwach mehr signal als noise
- insgesamt kann man davon ausgehen, dass miRNA:target Interaktionen hauptsächlich in den 3'UTR Bereichen der mRNA von Vertebraten vorkommen



- biologische Prozess Klassifikation der Vertebraten miRNA targets, vorhergesagt in der seed + t1A + m8 Analyse
- gezeigt sind Kategorien, deren targets eine signal:noise ratio von

Quellen

- **1) Prediction of Mammalian MicroRNA Targets**
Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, Christopher B Burge¹. *Cell*, Vol. 115 (7), December 26, 2003.
- **2) Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets**
Benjamin P Lewis, Christopher B Burge, David P Bartel². *Cell*, Vol. 120 (1), January 14, 2005.
- **Vortrag und Handout auf**
stinfwww.informatik.uni-leipzig.de/~mai03hfd