Vorhersage von miRNA targets in Säugetieren
- miRNA target prediction -
- am Beispiel von TargetScan –

Einleitung

- miRNAs ~22nt lang, Pflanzen- und Tierzelle
- 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 Gene
- 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

Algorithmus

- Ziel: Vorhersage von Zielgenen für konservierte miRNA in Wirbeltieren (Vertebraten)
- Prinzip: Identifizierung der mRNA, die eine konservierte Bindung mit "miRNA seed" eingeht
- kombiniert thermod. 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
- Input: - 1 miRNA, in mehrern Organsimen konserviert
  - Menge an orthologen[1] 3'UTR[2] Sequenzen dieser Organismen
- Schritte:
  1. Durchsuchen der 3'UTR Sequenzen der Organsimen mit Segmenten perfekter Watson Crick Basenkomplementarität zu den Basen 2-8 der miRNA
  2. Erweitern jedes seed matches mit zusätzlichen Basenpaaren zur miRNA soweit wie möglich in jede Richtung; G:U erlaubt, Abbruch bei Gaps
  3. "RNAfold" liefert optimierte Basenpaarung für das 3' Ende der miRNA zu den Basen der UTR, die direkt 5' jedes seed matches lokalisiert ist
  4. 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
  5. jede UTR erhält einen Z score $Zi=\sum_{k=1}^{n} e^{-Gk/T}$
     n=Anzahl der seed matches pro UTR; Gk ist die freie Energie der miRNA:target Interaktion (kcal/mol) für das kte target;T istGewichtungsfaktor, verbindet Z-Score und vorhergesagte freie Energie; gewichtet Affinität der Basenpaarung; Wert beeinflusst die relative Gewichtung der UTRs

[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)

6 sortieren der UTRs eines Organismus nach Z Score und Zuweisung eines Ranges Ri; Z und R dienen unter anderem dazu, daß eine Anzahl falsch Positiver targets optimal geschätzt werden können
7 wiederholt den beschriebenen Prozess für die Menge an UTRs in jedem Organismus
8 wählt als targets diejenigen Gene aus, für die Zi≥Zc und Ri ≤Rc; Zc = vorausgewählter Maximalwert für Z score; Rc = vorausgewählter Maximalwert für Rang; Rc, Zc und T sind frei bestimmbar; T und Z kommt keine thermodynamische Bedeutung zu sondern sie sind ein Mittel um vorhergesagte freie Energien zu gewichten und zu bewerten

Ergebnisse / Beobachtungen

- 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 (siehe 'Herstellung der zufällig permutierten Sequenzen')
- d.h. zur Kontrolle werden signal und noise miteinander verglichen
- im Endeffekt :
- 3.9 wahre targets pro miRNA in Säugern
- signal:noise ratio von 3.5
- je mehr Genome man in die Berechnungen integriert, 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 zwischen Säugern und Fischen konserviert
- 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

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 viel größer

Experimentelle Prozeduren

MicroRNA Datasets

- Mensch und Maus miRNAs Sequenzen wurden von Rfam heruntergeladen
- um innerhalb der Genome Homologe zu finden, wurde BLASTN und MIRscan

3' UTR Datasets

- 3' UTR Sequenzen für alle menschlichen Gene, und alle Maus, Ratte und Fugu Gene, die in Verbindung mit einem menschlichen Homolog standen, wurden mittels 'EnsMart'15.1 ermittelt
- annotierte 3'UTR Sequenzen waren nur für 45% der Ratten Gene und für keine der Fugu Gene verfügbar
- mehr als 14% der 3'UTR Ratten Sequenzen waren kürzer als 50 nt
- deswegen wurde jede annotierte 3'UTR mit 2kb von 3' flankierenden Sequenzen erweitert

Identifikation der miRNA TargetSites

- die 3' UTR Sequenzen wurden nach antisense matches der entsprechenden seed Region jeder miRNA durchsucht (Basen 2..8, beginnend vom 5' Ende)
- die Wahl eines 7nt langen seeds wird gestützt durch die Beobachtung, dass kürzere seeds eine geringere signal:noise ration zur Folge haben, und längere seeds, zu vergleichbaren signal:noise ratios die Anzahl der vorhergesagten targets reduzieren
- das Ändern der seed Länge hat einen entscheidenden Einfluss auf signal und noise
- Basenpaarungen zwischen miRNA und mRNA wurden soweit wie möglich in beide Richtungen verlängert, G:U Paare erlaubt, Gaps verboten
- die Basenpaarungsmuster des verbleibenden 3' Endes wurde mittels RNAfold vorhergesagt RNAfold wurde auf einer foldback Sequenz angewendet, die aus einem künstlichen stemloop besteht (5'-GGGCCCGGGULLLLLLACCCGGGCCC-3', wobei L eine anonyme ungepaarte loop Base ist, alle anderen Basen sind komplementär gepaart mit einer Base auf der gegenüberliegenden Seite des Stems)**,** der an den erweiterten seed match gehangen wurde

Parameter Optimierung

- die Trainingsdatensätze wurden mittels 40 zufällig ausgewählten miRNAs von Säugern und 27 zufällig ausgewählten miRNAs von Vertebraten konstruiert
- T wurde in 5er Schritten von 5 bis 25 erhöht
- Zc wurde in 0.5er Schritten von 1 bis 10 erhöht
- Rc wurde in 50er Schritten von 50 bis 1000 erhöht
- für die Parameter T=20, Zc=4.5. Rc=200 wurde eine optimale signal:noise ratio von 3.4:1 für den Säugetiertrainingssatz gefunden
- T=10, Zc=4.5 und Rc=350 gab eine optimale signal:noise ratio von 4.6:1 für den Vertebratentrainingssatz

Herstellung der zufällig permutierten Sequenzen

- für jede miRNA wurden zufällig permutierte Sequenzen erzeugt, die die gleiche Startbase, Länge und Basenkomposition hatten, bis 4 Sequenzen gefunden wurden, die sich von der Originalsequenz in weniger als 15% der folgenden Kriterien unterschieden:
- 1. E(SM) Markov Wahrscheinlichkeit 1. Ordnung des seed match
- 2. E(TM) Markov Wahrscheinlichkeit 1. Ordnung des Antisense des 3' Endes der miRNA
- 3. O(SM) beobachtete Anzahl der seed matches im UTR Datensatz
- 4. der vorhergesagten freien Energie eines seed:seed match Duplexes (RNAeval)

Neuerungen / Verbesserugen

- 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
- man fand heraus, 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
- 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)
- 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

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 Bartel ,2. Cell, Vol. 120 (1), January 14, 2005.

3.) TargetScan zum anwenden

- **Vortrag und Handout auf stinfwww.informatik.uni-leipzig.de/~mai03hfd**