Problemseminar Paper: Inference of miRNA targets using evolutionary conservation and pathway analysis

Andrej Aderhold

20. Juli 2007

Zusammenfassung *Der hier vorgestellte Artikel [1] handelt von der Anwendung des Bayes Theorems zur Schliessung auf miRNA Ziele. Für jede miRNA wird die Evolution von orthologen miRNA Zielen in einer Untergruppe von verwandten Organismen modelliert. Diese Methode schliesst die Verteilung von funktionalen miRNA Zielen, unabhängig für jeder miRNA, ein. Die vorgestellten Resultate lassen vermuten, dass diese Methode sehr gute Werte für die Vorhersage von miRNA Zielen liefert. Des weiteren werden die Vorhersagen mit einer Pathway Analyse kombiniert (KEGG), wobei für jede miRNA Funktionen zugeordnet werden.*

Target Suche: Seed Typen Die Suche wird ausschliesslich im mRNA 3'UTR durchgeführt, da dort die meisten Bindungsstellen vermutet werden. Als relevant werden die ersten 8 Nukleotide aus dem 5' Ende der mature miRNA Sequenz angenommen. Die Suche nach komplementären Sequenzen in Referenz und orthologen Genen wird unterschieden für verschiedene Seedtypen (Abb.1). Also die Länge des Komplemens und ggf. Nicht-Bindungen innerhalb oder am Rand der Komplements. Da die Stärke der Konservierung einer Sequenz Aufschluss über dessen Funktionalität geben kann, wurde ermittelt wie stark sich die Konservierung zwischen den seed Typen unterscheidet. Dafür wurde für jeden seed Typ und jeder Familie von Organismen (Säugetiere, Fliegen, Würmer, Fische) der Anteil von seed Sequenzen ermittelt der perfekt konserviert in allen Spezies der gleichen Familie ist. Dem gegenübergestellt wurde der Anteil perfekt konservierter Zufalls-Sequenzen die von Länge und Mismatches den seed Typen entsprechen. Das Verhältnis der beiden Anteile ist für jeden seed Typen und unterschiedlichen Familien in Abbildung 2 dargestellt. Wie vermutet zeigen die perfekt komplementär bindenden seed Typen die beste Konservierung (1-8, 1-7, 2-8). Die Autoren haben sich entschieden nur diese in ihre folgenden Berechnungen mit einzubeziehen.

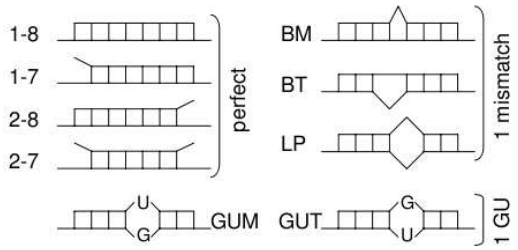

Abbildung 1: MiRNA seed Types

Algorithmus: Bayesianische phylogenetische miRNA Ziel Identifikation Gegeben sind ein Referenzorganismus und eine Gruppe von verwandten Arten. Mit gegebenen seed Typ t einer miRNA werden komplementäre Sequenzen in den 3' UTRs der Referenzspezies gesucht. Durch pairwise Alignment zwischen Referenz und jeweils einem Verwandten werden orthologe Sequenzen in den Verwandten ausgemacht. Man konstruiert einen binären Vektor $\vec{c}$ für jedes mutmassliche miRNA Ziel in der Referenz, wobei $c_i = 1$, falls Ziel ortholog in Spezies i und sonst

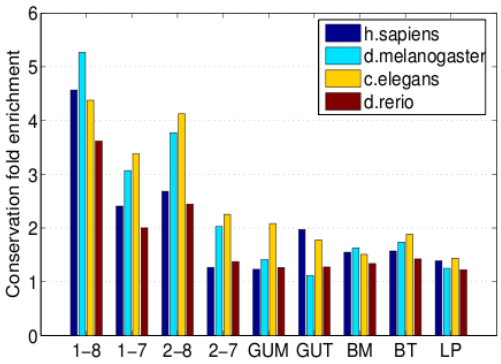

Abbildung 2: Grad der Konservierung bei verschiedenen seed Typen und Gruppen

$c_0 = 0$.
Ein Hintergrund Modell wird erstellt, bei dem in den 3' UTRs der Referenz zufällig Sequenzen ausgewählt werden, die in der Länge dem jeweiligen seed Typen t entsprechen. Orthologe Sequenzen werden in den Unterspezien gesucht. Ermittelt wird für jedes mögliche Konservierungsmuster $\vec{c}$ die relative Frequenz $p(\vec{c}|t, bg)$, mit denen es in den Alignments auftritt; nehme den Durchschnitt über alle seed Typen 1-8 match, 1-7 und 2-8. Das ist also die Wahrscheinlichkeit ein bestimmtes Konservierungsmuster zufällig im 3'UTR zu beobachten. Obwohl dies keinen absoluten Zufall darstellt, da aus dem 3'UTR ausgewählte Zufallssequenzen durchaus funktionale Elemente überlappen können und somit einem evolutionären Druck ausgesetzt sind. Erstelle ein Selektionsmuster, dass Auskunft über die Funktionalität der konservierten miRNA Ziele in den Unterspezien gibt. Die Selektion korrekt zu modellieren ist eigentlich unmöglich weil zu viele Faktoren von denen wir keine Kentniss haben Einfluss ausüben. Deswegen wird hier ein stark simplifizierter Ansatz angewendet. Wir nehmen an, dass ein miRNA Ziel funktional sei wenn eine Selektion darauf gewirkt hat die es konserviert lässt. Ein Ziel ist nicht-funktional wenn es nach dem Hintergrund Modell evolviert ist. Ein Selektionsmuster $\vec{s}$ ist ein binärer Vektor mit

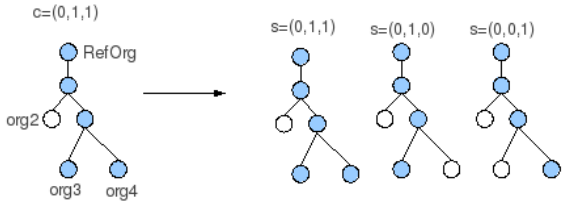

Abbildung 3: Konservierungsmuster und mögliche Selektionsmuster.

$s_i = 1$, falls ein miRNA Ziel in Spezies i funktional ist und $s_i = 0$ falls nicht. Sei $C(\vec{s})$ die Menge der Konservierungsmuster $\vec{c}$ die konsistent mit dem Selektionsmuster $\vec{s}$ ist , sodass $\forall \vec{c} \in C$ gilt $c_i = 1$ für alle i für welche $s_i = 1$. Also wähle alle Konservierungsmuster aus in den Spezies vorkommen (konserviert sind) in denen auch eine Selektion festgelegt ist. Ist ein Ziel in einem Organismus unter Selektion dann muss er auch zwangsweise konserviert sein! Aber ein konserviertes Ziel muss nicht zwangsweise und Selektion stehen. Damit lässt sich für jedes Konservierungsmuste eine Menge von möglichen Selektionsmustern zuordnen 3. Die Wahrscheinlichkeit

$$p(\vec{c}|t, \vec{s}) = \frac{p(\vec{c}|t, bg)}{\sum_{\vec{c'} \in C(\vec{s})} p(\vec{c}|t, bg)}$$

gibt die Möglichkeit an $\vec{(c)}$ zu beobachten unter gegebenen Selektionsmuster $\vec{s}$. Weiter wird die Verteilung $p(\vec{s})$ für jede miRNA genutzt um die Wahrscheinlichkeit festzulegen mit der ein mutmassliches miRNA Ziel in der Referenz unter Selektion in allen Unterspezien ist, bei denen $s_i = 1$ definiert ist. Um den Einfluss der Konservierung der miRNA zu Berücksichtigen wird überprüft ob es Unterspezien i gibt bei denen $s_i = 1$ aber das miRNA Gen nicht konserviert ist. Ist dies der Fall so wird für diese Selektionsmuster $p(\vec{s} = 0)$ gesetzt. Das signalisiert, dass ein Selektionmuster unbrauchbar wird falls ein miRNA Ziel unter Selektions steht aber die miRNA nicht vorkommt (weswegen kein Selektionsdruck da ist). Schliesslich muss die Wahrscheinlichkeitsverteilung von $p(\vec{(s)})$ gefunden wer-

den, die am nähesten die beobachteten Daten erklärt. Mit

$$p(\vec{c}|t) = \sum_{\vec{s}\in S} p(\vec{c}|t,\vec{s})p(\vec{s})$$

wird die Wahrscheinlichkeit ermittelt $\vec{c}$ zu beobachten mit seed Typ t, wobei die Menge S alle Selektionsmuster enthält, die konsistent mit dem miRNA Gen Konservierungsmuster sind. Sei $n(\vec{c},t)$ die Anzahl vorkommender mutmasslicher miRNA Ziele mit Konservierungsmuster $\vec{c}$ und seed Typ t in der Referenz. Die likelihood Funktion L, gegeben dem $n(\vec{c},t)$ ist dann

$$L = \prod_{\vec{c},t} p(\vec{c}|t)^{n(\vec{c},t)}$$

Ermittle $p(\vec{s})$ (Likehood Parametervektor) durch Maximierung von L. Damit die Evolution des Selektionsdrucks entlang des phylogenetischen Baumes besser nachgebildet werden kann und aufgrund der hohen Anzahl von Parametern (2^n für n Unterspezies), die zu 'overfitting' führen können schlagen die Autoren neue Parameter für $p(\vec{s})$ vor: Man nehme den phylogenetischen Baum und ordne jeder Verzweigung drei Wahrscheinlichkeiten $p_w(k)$ zu, wobei $w \in \{01, 10, 11\}$ mit $w = 10$ für den linken Abkömmling steht, w=11 für beide Abkömmlinge und $w = 01$ für den Rechten (Abb.4). Die Variable k identifiziert die Verzweigung (innerer Knoten) innerhalb des Baumes. Die Wahrscheinlichkeit gibt nun an wie gross der Anteil an funktionalen Zielen ist auf die Selektionsdruck ausgeübt wird an jedem Knoten k für die Linke, Rechte und beide Untergruppen. Sei zudem p der Anteil an funktionalen miRNA Zielen in der Referenz die mindestens ein orthologes Ziel besitzen, das unter Selektion steht. Dieser Anteil p gibt somit die untere Grenze an mit der zu erwarten ist, dass funktionale Regionen in den Unterspezien auftreten. Er wird berechnet indem die miRNA Ziele in der Referenz die mindestens ein konserviertes orthologen Ziel in einem anderen Organismus besitzen zum Verhältnis mit allen gefundenen Zielen in der Referenz gesetzt werden.

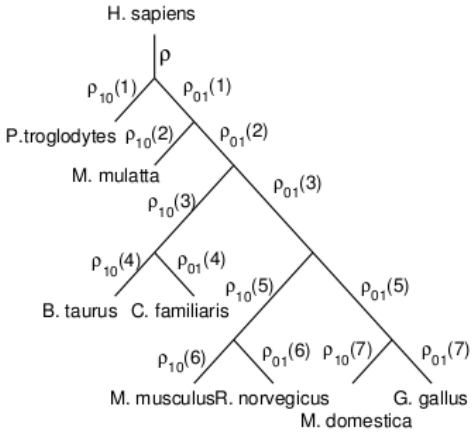

Abbildung 4: Phylogenetischer Baum mit Parametern: Wahrscheinlichkeiten an jedem Knoten funktionale Elemente beizubehalten

In Verbindung mit der Wahrscheinlichkeitsverteilung der Selektionsmuster, dient sie als untere Schranke mit der festgelegt wird, wie hoch mindestens die Chance ist ein beliebiges Selektionsmuster mit gegebenen Konservierungsmuster zu beobachten. Als Beispiel können für die in Abbildung 3 dargestellten Selektionsmuster folgende Wahrscheinlichkeiten festgelegt werden: $p(0,1,1) = pp_{01}(1)p_{11}(2)$, $p(0,1,0) = pp_{01}(1)p_{10}(2)$, $p(0,0,1) = pp_{01}(1)p_{01}(2)$. Um nun $p(\vec{s})$ zu erhalten müssen alle $p_w(k)$ ermittelt werden, indem das likehood der Wahrscheinlichkeitsverteilung maximiert wird, gegeben den Daten. Sei $\delta(\vec{s},\omega,k)$ eine Indikatorfunktion, die $\delta(\vec{s},\omega,k) = 1$ gesetzt wird falls $p_w(k)$ in $p(\vec{s})$ auftritt und $\delta(\vec{s},\omega,k) = 0$ falls nicht. Dadurch ergeben sich $\forall\omega, k$ die Ableitungen

$$\frac{dp(\vec{s})}{dp_\omega(k)} = \delta(\vec{s},\omega,k)\frac{p(\vec{s})}{p_w(k)}$$

Mit dieser Formel und durch Anwenden der Expectation Maximization Prozedur (EM) können $p_w(k)$ so approximiert werden, dass L Maximal. Dafür sei ein $X_\omega(k)$ definiert als

$$X_\omega(k) = \sum_{\vec{c},t} n(\vec{c},t)[\sum_{\vec{s}\in S}\delta(\vec{s},\omega,k)\frac{p(\vec{c}|t,\vec{s})p(\vec{s})}{\sum_{\vec{\sigma}\in S} p(\vec{c}|t,\vec{\sigma})p(\vec{\sigma})}]$$

dann ist die EM Update Formel gegeben durch

$$p_\omega(k) = \frac{X_\omega(k)}{\sum_{\tilde{\omega}\in\{01,10,11\}} X_{\tilde{\omega}(k)}}.$$

Diese Funktion muss iteriert werden um optimale $p_\omega(k)$ zu erhalten.
Die Bayes Formel kann nun für jedes Ziel, gegeben seinem Konservierungsmuster, angewendet werden. Sei zusätzlich noch $\vec{0}$ das Selektionmuster, das keine Selektion hat, also $\vec{0} = (0, .., 0)$, dann

$$p(\vec{s} \neq \vec{0}|t, \vec{c}) = 1 - \frac{p(\vec{c}|t, bg)(1-p)}{\sum_{\vec{s}\in S} p(\vec{c}|t, \vec{s})p(\vec{s})}$$

Der Parameter $(1-p)$ ist die a priori Wahrscheinlichkeit der Nicht-Selektion ($p(\vec{0})$).

Bayes Methode: Resultat Zum Vergleich mit anderen Methoden wurden die Vorhersagen an 120 experimentell nachgewiesenen mRNA-miRNA Interaktionen im Fisch getestet. Für die hier vorgestellte Methode wurden verschiedene posterior Wahrscheinlichkeits cut-offs festgelegt aus denen sich verschiedene Mengen von miRNA Ziel Vorhersagen ergeben. Diese sind als schwarze Linie in Abb. 5 dargestellt. Die Abbildung lässt erkennen, dass die Methode so gut wie die Akkuratesten (Grün et al.[2], Stark et al. [3]) ist und dabei eine hohe Spezifität bei einer gleichzeitig hohen Sensivität aufweist. Desweiteren wurde untersucht inwieweit sich die vorhergesagten miRNA Ziele bei den Fliegen mit denen von Grün et al. [2] und Stark et al. [4] überlappen. Es wurde festgestellt, dass die Unterschiede zwischen den miRNAs sehr gross sind. Für die miRNAs des Zuckermais z.B. besteht eine Übereinstimmung von 76% (Grün) und 86% (Stark); die Diskrepanz ist stärker bei mir-1: 75% (Grün) und 70% (Stark) und für mir-281: 50% (Grün) und 38% (Stark). Das bedeutet, dass mindestens die Hälfte der vorhergesagten Ziele nicht von den anderen gefunden wurden. Um einen fundierten Vergleich zu führen bedarf es aber mehr experimenteller Daten von

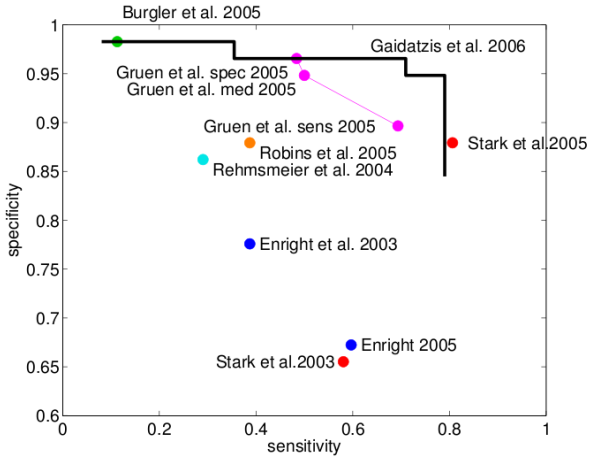

Abbildung 5: Vergleich zu anderen Methoden: Spezifität vs. Sensivität

miRNA-mRNA Interaktionen von denen es zu wenige gibt.

miRNA Funktion aus der Pathway Analyse Um herausfinden in welchen biochemischen Pfaden die einzelnen miRNAs eine Rolle spielen könnten wurden die Verhersagen mit der KEGG Datenbank [5] kombiniert in der viele Gene einer metabolischen Funktion zugeordnet sind. Als cut-off für die posterior Wahrscheinlichkeit wurde ≥ 0.5 festgelegt. Um die statistische Signifikanz der Zuordnung eines Pfades zu einer miRNA zu ermitteln wurde ein log likelihood ratio für die beobachteten Daten errechnet. Dieser setzt sich aus einem unabhängigen und abhängigen Modell zusammen. Beim unabhängigen Modell ist die Wahrscheinlichkeit der Interaktion einer mRNA mit einer miRNA unabhängig von der Zuordnung der mRNA zu einem Pfad. Beim abhängigen Modell spielt die Zuordnung der mRNA eine Rolle. Der Ratio aus den likelihoods gibt den Grad der Wahrscheinlichkeit an, dass miRNA und ein biochemischer Pfad assoziert sind. Der Wert kann entweder positiv sein (miRNA Ziele sind verstärkt im Pfad) oder negativ (miRNA Ziele sind unterrepräsentiert). Diese Methode zeigt die erwar-

teten Assoziationen für miRNAs die in spezifischen Zellen aktiv sind und für miRNAs deren Gen Ziele schon bekannt sind. Wie schon in [3] gezeigt sind weitverbreitete Gene, die grundlegende metabolische Funktionen innenhalten selten Ziele von miRNAs. Ziele sind v.a. Gene der Transcriptionsregulation, Interzellulären Kommunikation, Zell Wachstum, Tod und Entwicklung. Besonders verbreitet scheint die Involvierung von miRNAs in die Entwicklung des Nerven Systems zu sein: 19 von 119 miRNA spielen eine Rolle bei der Axon Führung. In dem Artikel wurden dazu einige miRNAs als Beispiele herangezogen.

Literatur

[1] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M: **Inference of miRNA targets using evolutionary conservation and pathway analysis**. *BMC Bioinformatics* 2007, **8**:69.

[2] Grün D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N: **micro-RNA target predictions across seven Drosophila species and comparison to mammalian targets**. *PloS Comput Biol* 2005, **1**:e13.

[3] Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM: **Animal Micro-RNAs conver robustness to gene expression and have a significant impact on 3'UTR evolution**. *Cell* 2005, **123**:1133–1146.

[4] Stark A, Brennecke J, Russell RB, Cohen SM: **Identification of Drosophila miRNA targets**. *PLoS Biol* 2003, **1**(3):e60.

[5] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Research* 2006, **34**:D354–7.