comparative structure prediction of ncRNA molecules

using a non Sankoff approach

Sebastian Bartschat

01. Februar 2008

Inhalt

Was sind ncRNA's ?

ncRNA

non coding RNA-Moleküle sind DNA-Sequenzen, die transkripiert, aber nicht translatiert werden.

- vielfältigste regulatorische Funktion, z.B.
  1. gene silencing (miRNA)
  2. Replikation (Telomerase RNA)
  3. rRNA-Reifung und Modifikation (snoRNA)
- schwer zu identifizieren

bedeutende Struktur

- ausgeprägte Sekundärstruktur, die oft funktionsbedingt ist
  z.B. autokatalytisches Gruppe I Intron bei Prokaryoten
- wichtig die Struktur vorherzusagen, um so Funktionen und Komplexität besser zu verstehen

Matthews Correlation Coefficient (MCC)

$$\text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

dabei gilt: $-1 \leq \text{MCC} \leq 1$

Strukturvorhersage

Minimal Free Energy (MFE)

- benötigt viele Parameter (stem energy, loops, Temperatur . . .)
- Struktur mit geringster freier Energie ist oft nicht die Richtige
- Liste von suboptimalen Faltungen durch RNAsubopt

Strukturvorhersage

comparative prediction

Sekundästrukturen sind innerhalb von ncRNA Familien relativ stark konserviert (im Gegensatz zu ihrer Sequenz).

Annahme:	gemeinsame Struktur von verwandten ncRNA-Sequenzen entspricht eher der natürlichen Struktur
Möglichkeit:	Entdeckung neuer ncRNA-Familien

Strukturvorhersage

Sankoff-Ansatz

- gleichzeitiges Durchführen von Sequenzalignment und Strukturvorhersage von multiplen RNA Sequenzen
- sehr Speicher- und CPU-belastend
- viele Programme verfolgen diesen Ansatz und versuchen Ressourcen zu schonen
  z.B. stemloc, Pmcomp usw.

Strukturvorhersage

nicht Sankoff basierte Ansätze

RNAspa und RNAcast

- RNAcast kümmert sich nicht um Alignment der gegebenen Sequenzen
- versucht lediglich die Sekundärstruktur vorherzusagen
- arbeitet mit *shapes*

- RNAcast kümmert sich nicht um Alignment der gegebenen Sequenzen
- versucht lediglich die Sekundärstruktur vorherzusagen
- arbeitet mit *shapes*

shape

Familie von Sekundärstrukturen, die einen gemeinsamen Grundaufbau besitzen.

Grundzüge

1. zu jeder Sequenz werden alle (innerhalb einer gewissen Grenze) suboptimalen Faltungen brechnet $\rightarrow$ ergibt den sogenannten Faltungsraum

Grundzüge

1. zu jeder Sequenz werden alle (innerhalb einer gewissen Grenze) suboptimalen Faltungen brechnet → ergibt den sogenannten Faltungsraum
2. Berechnung der jeweiligen shapes → reduziert den zu betrachtenden Raum

Grundzüge

1. zu jeder Sequenz werden alle (innerhalb einer gewissen Grenze) suboptimalen Faltungen brechnet → ergibt den sogenannten Faltungsraum
2. Berechnung der jeweiligen shapes → reduziert den zu betrachtenden Raum
3. Suche nach *common shapes*; d.h. shapes, die in allen Faltungsräumen vorkommen

Grundzüge

1. zu jeder Sequenz werden alle (innerhalb einer gewissen Grenze) suboptimalen Faltungen brechnet $\rightarrow$ ergibt den sogenannten Faltungsraum
2. Berechnung der jeweiligen shapes $\rightarrow$ reduziert den zu betrachtenden Raum
3. Suche nach *common shapes*; d.h. shapes, die in allen Faltungsräumen vorkommen
4. Bestimme *consensus shape* zwischen allen *common shapes*

RNA shape analysis

- jede Sequenz s besitzt Faltungsraum $\mathcal{F}(s)$; zu jeder Faltung $x \in \mathcal{F}(s)$ kann die freie Energie $E(x)$ berechnet werden
- $mfe(s)$ ist diejenige Struktur $x \in \mathcal{F}(s)$, für die $E(x)$ minimal ist
- $\mathcal{S}$ ... string-like domain von Strukturen;
  $\mathcal{P}$ ... string-like domain für shapes
- π ist eine Mappingfunktion von $\mathcal{S}$ nach $\mathcal{P}$

RNA shape analysis

- Menge aller shapes: $\mathcal{P}(s) = \{ \pi(x) \mid x \in \mathcal{F}(s) \}$
- Klasse aller *p-shaped*-Strukturen:

$$\mathcal{F}(s \mid p) = \{ x \mid x \in \mathcal{F}(s), \pi(x) = p \}$$

- *shrep* $\hat{p}$ ist diejenige Struktur eines *shapes* p mit der kleinsten freien Energie

Vorteile

- gibt guten Überblick über den Faltungsraum $\longrightarrow$ Reduzierung auf wenige *shapes*
- generischer Ansatz, da *shapes* je nach Wahl von π mehr oder weniger abstrakt sind

Mappingregeln

$$
\begin{array}{ll}
\pi_5(.) = \varepsilon & \varrho_5(.) = \varepsilon \\
\pi_5(.s) = \pi_5(s) & \varrho_5(.s) = \varrho_5(s) \\
\pi_5(s.) = \pi_5(s) & \varrho_5(s.) = \varrho_5(s) \\
\pi_5((s)) = [\varrho_5(s)] & \varrho_5((s)) = \varrho_5(s) \\
\pi_5((s)s') = [\varrho_5(s)]\pi_5(s') & \varrho_5((s)s') = \pi_5((s)s')
\end{array}
$$

$$
\begin{array}{ll}
\pi_3(.) = \varepsilon & \varrho_3((s)) = \varrho_3(s) \\
\pi_3(.s) = \pi_3(s) & \varrho_3(s) = \pi_3(s) \text{ In all other cases} \\
\pi_3(s.) = \pi_3(s) & \\
\pi_3((s)) = [\varrho_3(s)] & \\
\pi_3((s)s') = [\varrho_3(s)]\pi_3(s') &
\end{array}
$$

Struktur : '(((...((..(((...)))))).((...))..)))'

Level 5: '[][]]'

Level 3: '[[[]][]]'

Definitionen

`Definition 1`

p ist *common shape* $\Leftrightarrow\ p \in \bigcap_{i=1}^{k} \mathcal{P}(s_i)$

`Definition 2`

consensus shape für Sequenzen $\{s_1 \dots s_k\}$ ist derjenige *common shape* p mit minimalem rank $(\hat{p}_1 \dots \hat{p}_k)$

Algorithmus

step 1 `INPUT:` Sequenzen $\{s_1, \ldots, s_k\}$ und Schranke $\mathcal{R}$
RNAshapes berechnet zu jeder Sequenz $\mathcal{F}(s_i)$ und $\mathcal{P}(s_i)$

Algorithmus

step 1 `INPUT`: Sequenzen $\{s_1, \ldots, s_k\}$ und Schranke $\mathcal{R}$
RNAshapes berechnet zu jeder Sequenz $\mathcal{F}(s_i)$ und $\mathcal{P}(s_i)$

step 2 Berechne alle *l common shapes* und ihre *shreps* :

$$[\,(p_1, [\,\hat{p}_1^1, \ldots, \hat{p}_k^1\,]), \ldots, (p_l, [\,\hat{p}_1^l, \ldots, \hat{p}_k^l\,])]$$

Algorithmus

step 1 `INPUT:` Sequenzen $\{s_1, \ldots, s_k\}$ und Schranke $\mathcal{R}$
RNAshapes berechnet zu jeder Sequenz $\mathcal{F}(s_i)$ und $\mathcal{P}(s_i)$

step 2 Berechne alle *l common shapes* und ihre *shreps* :

$$[\,(p_1, [\,\hat{p}_1^1, \ldots, \hat{p}_k^1\,]), \ldots, (p_l, [\,\hat{p}_1^l, \ldots, \hat{p}_k^l\,])]$$

step 3 Bestimme zwischen allen l *common shapes* mittels einer geeigneten scoring-Funktion den *consensus shape*
`OUTPUT:` $(p\,, [\,\hat{p}_1, \ldots, \hat{p}_k\,])$

Ergebnisse

```
Shape: [[[]][][]]        Score: -223.50
CCUUUGCAGGCAGCGGAAAUCCCCACCUGGUAACAGGUGCCUCUGCGGCCAAAAGCCACGUGUAUAAGAUACACCUGCAAAGG
((((((((((..(((((......((((((....))))))..))))))((.....)))..((((.....))))..))))))))))   (-34.10) R = 2
CCUUUGCAGGCAGCGGAAUCCCCCACCUGGUGACAGGUGCCUCUGCGGCCGAAAGCCACGUGUGUAAGACACACCUGCAAAGG
((((((((((..(((((......((((((....))))))..))))))((.....)))..((((.....))))..))))))))))   (-39.10) R = 2
GCACGCAAGCCGCGGGAACUCCCCCUUGGUAACAAGGACCCGCGGGGCCGAAAGCCACGUUCUCUGAACCUUGCGUGU
((((((((((((((.......((((....))))).))))))))((.....)))..((((...))))))))))))              (-34.10) R = 2
GCAUGAUGGCUGUGGGAACUCCCCCUUGGUAACAAGGACCCACGGGGCCAAAAGCCACGUCCUCACGGACCCAUCAUGC
((((((((((((((.......((((....))))).))))))))((.....)))..((((....))))))))))))             (-34.70) R = 3
GCAUGACGGCCGUGGGAACUCCUCCUUGGUAACAAGGACCCACGGGGCCAAAAGCCACGCCCACACGGGCCCGUCAUGU
((((((((((((((.......((((....))))))))))))))((.....)))..((((....))))))))))))             (-41.90) R = 1
GCAUGUUGGCCGUGGGAACACCUCCUUGGUAACAAGGACCCACGGGGCCGAAAGCCAUGUCCUAACGGACCCAACAUGU
((((((((((((((.......((((....))))))))))))))((.....)))..((((....))))))))))))             (-39.60) R = 1
```

Voraussetzungen

Grundlage für den Algorithmus ist RNAsubopt (Vienna Package): liefert eine Liste von suboptimalen Faltungen (bzgl. einer gewissen Schranke) für eine bestimmte Sequenz.

Algorithmus I

INPUT: n unalignierte und randomisierte Sequenzen S_1 , S_2 , . . . , S_n

step 1: RNAsubopt liefert zu jeder Struktur v mögliche Faltungen:

$$S_i^1 , S_i^2 , \ldots , S_i^v$$

die laut Annahme die optimale Struktur enthalten

Algorithmus II

step 2: alle S_i^j bilden nun einen Knoten im Graphen; die v verschiedenen Strukturen zu einer Sequenz S_i bilden dabei eine Schicht; wobei Schicht von S_i über der von S_{i+1} liegt, sodass alle n Schichten übereinander gestapelt sind

Algorithmus II

step 2: alle S_i^j bilden nun einen Knoten im Graphen; die v verschiedenen Strukturen zu einer Sequenz S_i bilden dabei eine Schicht; wobei Schicht von S_i über der von S_{i+1} liegt, sodass alle n Schichten übereinander gestapelt sind

jeder Knoten der Schicht S_i ist über gerichtete Kanten mit allen Knoten der Schicht S_{i+1} verknüpft

das Gewicht einer Kante repräsentiert die Ähnlichkeit der beiden adjazenten Knoten

Algorithmus III

step 3: mittels Breitensuche wird der kürzeste Pfad von Schicht 1 nach Schicht n bestimmt
Kosten ergeben sich aus der Summe der einzelnen Kantengewichte
$\longrightarrow$ dabei gilt: je kürzer der Pfad, desto ähnlicher die Strukturen

Algorithmus IV

step 4: Wiederhole die Schritte 1 - 3 einige Male
Bestimme zwischen allen kürzesten Pfaden denjenigen mit dem kleinsten sum-of-pair score aller $\frac{n(n+1)}{2}$ Knotenpaare

OUTPUT: n Strukturen, die sich untereinander so ähnlich wie möglich sind

Flaschenhals:

- Aufbau des Layers S_{i+1} erfordert V^2 Strukturvergleiche mit Layer S_i
- Wie schnell sind diese möglich und wie oft sind sie nötig?
- string Edit-Distance (ED) in $\mathcal{O}(N^2)$ (Needleman-Wunsch)

Flaschenhals:

- Aufbau des Layers S_{i+1} erfordert V^2 Strukturvergleiche mit Layer S_i
- Wie schnell sind diese möglich und wie oft sind sie nötig?
- string Edit-Distance (ED) in $\mathcal{O}(N^2)$ (Needleman-Wunsch)

Beschleunigung:

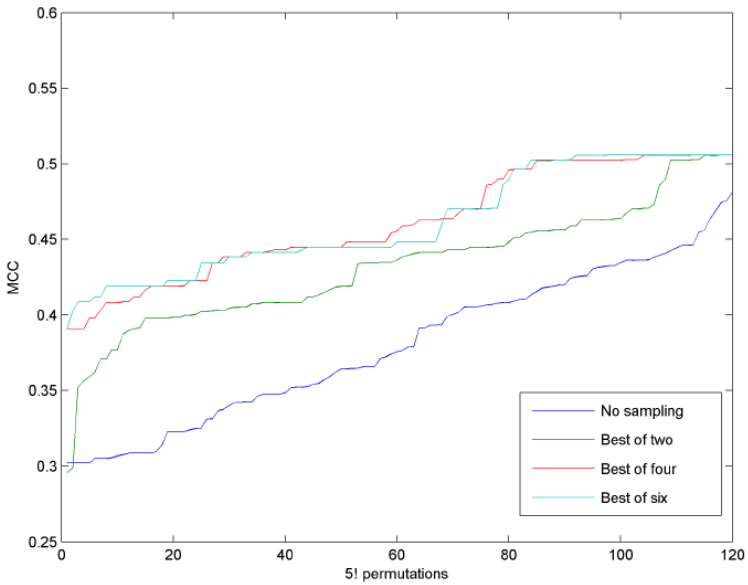
- Erlaube nur K Mismatches $\rightarrow$ führt zu $\mathcal{O}(KN)$
- Berechne nur "sinnvolle" Kanten
- bis zu 25% Zeitersparnis

Laufzeit:

$$\mathcal{O}(NL^3 + SNV^2L^2 + SN^2L^2)$$

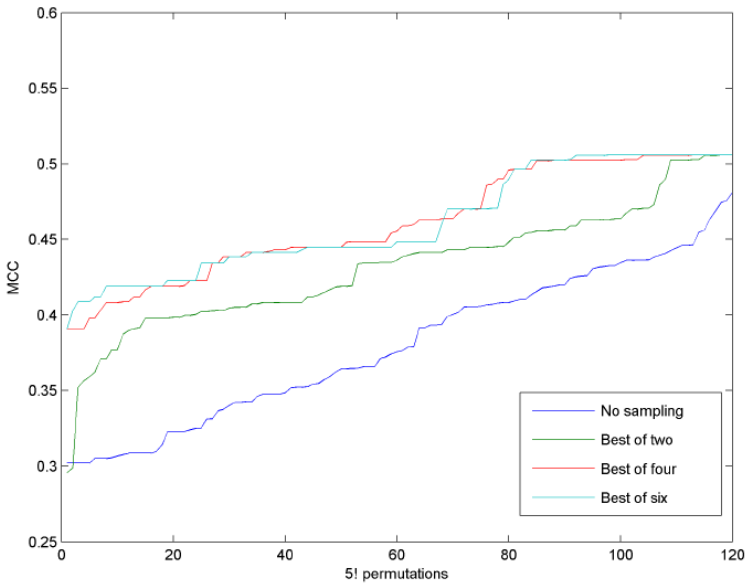
- N . . . Anzahl der Sequenzen
- L . . . Sequenzlänge
- V . . . Anzahl der suboptimalen Faltungen
- S . . . Anzahl der Durchläufe

Einfluss der Parameter

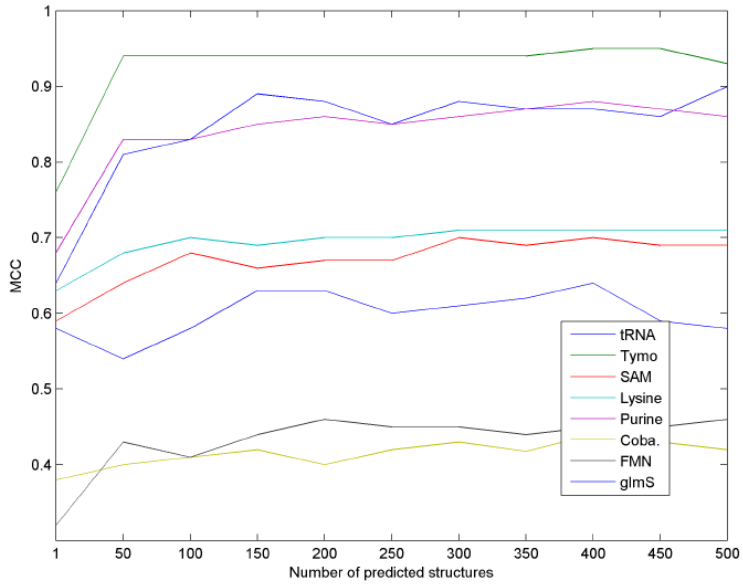


Einfluss des Samplespace...

Einfluss der Parameter

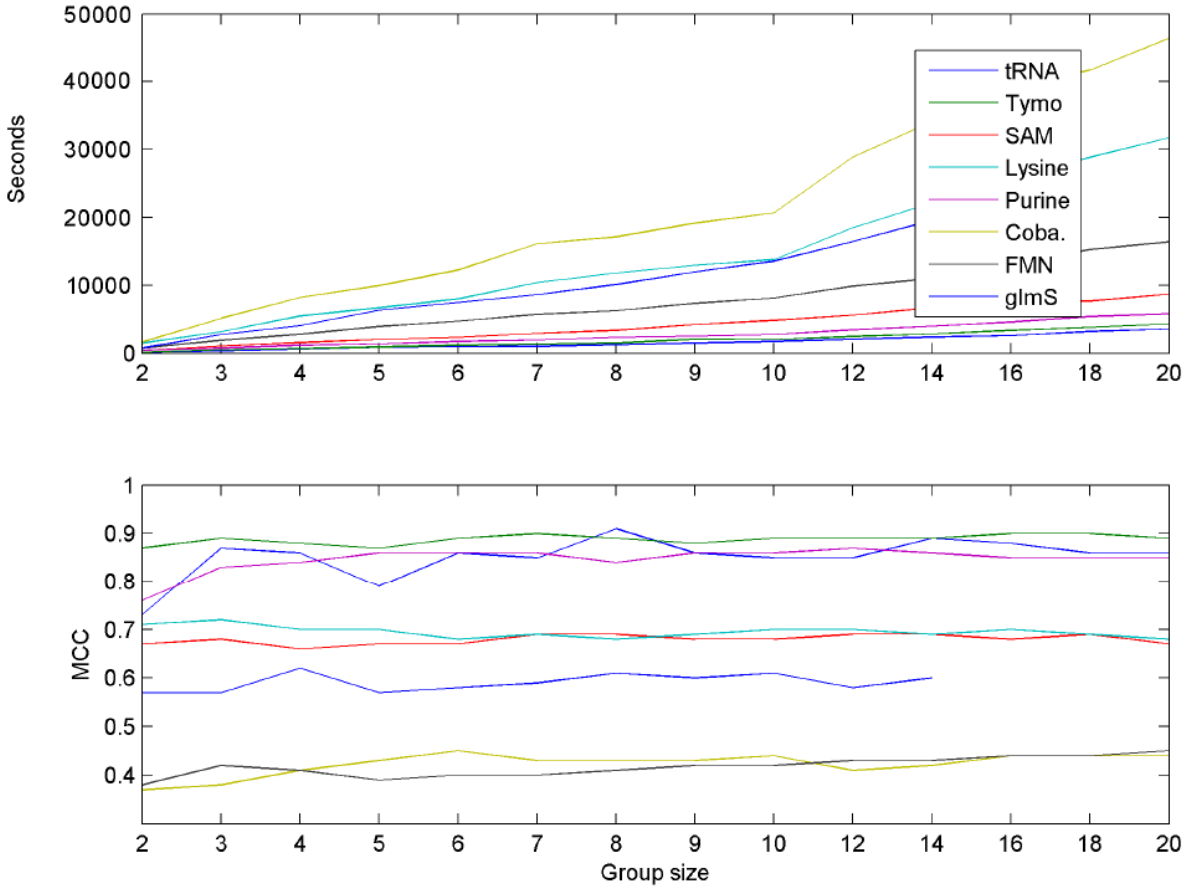


Einfluss des Samplespace...



Einfluss der suboptimalen Strukturen...

Einfluss der Parameter



Einfluss der Sequenzanzahl...

Dataset

lin4 miRNA	*microRNA* zur Regulation von Tranksriptionsvorgängen
tRNA	transfer RNA für Aminosäuren
5S rRNA	Bestandteil der großen ribosomalen Untereinhet
SRP RNA	dient der Modifizierung von Proteinen
IRES	Konformation führt zu Translationsstart
purine RS	Regulation der Proteinbiosynthese
SAM RS	Regulation der Termination der Transkription
snRNA	U1, U2, U12 erfüllen verschiedene Regulationsmechanismen

Geschwindigkeit

Family	Average run-time per sequence in seconds				
	RNAspa	StemLoc	pmMulti	FoldAlignM	RNAcast
tRNA	**13**	47	21	549	<1
Tymo	**8**	70	12	93	<1
SAM	**13**	3	18	248	<1
Lysine	**20**	11	69	1076	<1
Purine	**84**	339	135	1518	<1
Cobalamin	**133**	105	no data	no data	<1
FMN	**56**	6	141	5042	<1
glmS	**58**	60	no data	no data	no data

Genauigkeit

Family	Avg. sequence identity	MCC				
		RNAspa	RNAcast			
			-t 5 –c 10 (default)	-t 5 –c 20	-t 3 –c 10	-t 3 –c 20
lin4	66%	**0.86**	0.73	0.73	no data	0.72
tRNA	50%	**0.81**	0.42	0.42	no data	0.65
5S RNA	59%	**0.58**	0.62	0.62	no data	0.84
U2	78%	**0.74**	0.72	0.72	0.69	0.69
S box riboswitch	67%	**0.75**	0.69	0.69	0.78	0.78
SRP RNA	52%	**0.95**	0.99	0.99	0.94	0.94
IRES	66%	**0.97**	0.93	0.93	0.90	0.90
Purine riboswitch	56%	**0.93**	0.78	0.78	0.81	0.81
U1	65%	**0.67**	0.72	0.72	no data	0.76
U12	83%	**0.76**	0.50	0.50	no data	0.68

Robustheit

Family	MCC	
	RNAspa	RNAcast
lin4	**0.86**	no data
tRNA	**0.80**	0.42
5S RNA	**0.51**	no data
U2	**0.72**	no data
S box riboswitch	**0.77**	0.82
SRP RNA	**0.93**	0.78
IRES	**1.00**	no data
Purine riboswitch	**0.90**	0.78
U1	**0.61**	no data
U12	**0.83**	no data

Limitationen von RNAcast

- Wahl der richtigen Parameter
- findet nicht immer einen *common shape*
- Verunreinigungen durch unähnliche Sequenzen

RNAcast:

→ singnifikant schneller als alle anderen Programme

→ probleme mit kontaminierten Datensätzen

RNAcast:

→ singnifikant schneller als alle anderen Programme

→ probleme mit kontaminierten Datensätzen

RNAspa:

→ gute Genauigkeit in der Strukturvorhersage

→ sehr robuster Algorithmus

Literatur


Yair Horesh, Tirza Doniger, Shulamit Michaeli, Ron Unger *RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules*, Bioinformatics 2007 Jun 7; 8(366)


Robert Griegerich, Jens Reeder *Consensus Shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction*, Bioinformatics: Vol.21 no. 17 2005, pages 3519-3523


Stefan Wuchty, Walter Fontana, Ivo L. Hofacker, Peter Schuster *Complete Suboptimal Folding of RNA and the Stability of Secondary Structure* Biopolymers: Vol 49, 145-165 (1999)