Johann Wolfgang Goethe Universität Frankfurt am Main
Fachbereich 15: Biologie und Informatik
Institut für Informatik
Juniorprofessur für Bioinformatik
Dr. Dirk Metzler

Seminar: Aktuelle Themen der Bioinformatik

Arbeit von: Volker Hähnke

Ye Ding, Charles E. Lawrence: „A statistical sampling algorithm for RNA secondary structure prediction"

Elena Rivas, Sean R. Eddy: „Noncoding RNA gene detection using comparative sequence analysis"

Abstract

Die RNA ist für alle Lebewesen essentiell und beteiligt an einer Vielzahl von biochemischen Prozessen. Entscheidend für ihre Funktion ist dabei die Sekundärstruktur, die aus einer Kombination von wenigen immer wiederkehrenden Strukturmotiven besteht.
Der erste Teil dieser Ausarbeitung wird sich daher damit befassen, für eine gegebene Ribonukleotidsequenz eine zugehörige Sekundärstruktur (ohne Berücksichtigung eventuell möglicher Pseudoknoten) vorherzusagen. Der vorgestellte von Ye Ding und Charles Lawrence entwickelte Algorithmus bedient sich dabei der Gibbs-Boltzmann-Verteilung der möglichen Sekundärstrukturen von Teilsequenzen, die ausgehend von der freien Energie der einzelnen Sekundärstrukturen des zu faltenden Sequenzabschnittes gebildet wird. Im ersten Schritt des Algorithmus werden mittels dynamischer Programmierung die dazu notwendigen Zustandssummen der Gibbs-Boltzmann-Verteilung berechnet. Im zweiten Schritt wird der zu faltenden Sequenz mit den von der Gibbs-Boltzmann-Verteilung gelieferten Wahrscheinlichkeiten für Sekundärstrukturen für Teilsequenzen der Gesamtsequenz eine Kombination von Sekundärstrukturen zugewiesen.
Der zweite Teil dieser Ausarbeitung beschäftigt sich damit, auf der Ebene der DNA Bereiche zu identifizieren, die für RNA kodieren, die nicht translatiert wird, also wie zum Beispiel die Ribozyme eine determinierte Funktion hat. Diese RNA wird als funktionelle oder auch nicht kodierende RNA bezeichnet. Das vorgestellte, von Elena Rivas und Sean Eddy entwickelte, Analyseverfahren beruht auf dem Alignieren von homologen Nukleotidsequenzen. Durch drei Modelle, die jeweils die mögliche Evolution für Proteine kodierender, für funktionelle RNA kodierender oder sonstiger Sequenzabschnitte beschreiben, wird versucht, die im Alignment beobachteten Mutationsmuster zu erklären. Durch die Verwendung bereits bestehender Algorithmen zur Analyse der verwendeten Grundmodelle (Markov-Modelle und stochastisch kontextfreie Grammatiken) wird die Bayes'sche Posteriori-Wahrscheinlichkeit jedes Modells berechnet, die angibt, wie wahrscheinlich es ist, dass das betreffende Modell dem Eingabe-Alignment zuzuordnen ist. Hat das die funktionelle RNA repräsentierende Modell die höchste Wahrscheinlichkeit, werden die analysierten Sequenzabschnitte als Kandidaten für funktionelle RNA eingestuft.

A statistical sampling algorithm for RNA secondary structure prediction (Ye Ding, Charles Lawrence)

Motivation

RNA-Moleküle sind an einer Vielzahl von biologischen Prozessen beteiligt. Sie können als Ribozyme katalytische Funktion haben, sich selbst in der Translation regulieren oder an anderen Wechselwirkungen zwischen Nukleinsäuren bzw. Proteinen partizipieren. Maßgeblich verantwortlich für ihre Funktion ist hierbei die von der Ribonukleotidkette ausgebildete Sekundärstruktur, die bei genauerer Betrachtung eine zum Teil ineinander verschachtelte Kombination weniger, immer wiederkehrender, Strukturmotive ist. Abbildung 1 zeigt die Grundmotive, die in dem von Ding und Lawrence entwickelten Verfahren als die Grundbausteine der Sekundärstruktur einer RNA-Sequenz betrachtet werden.

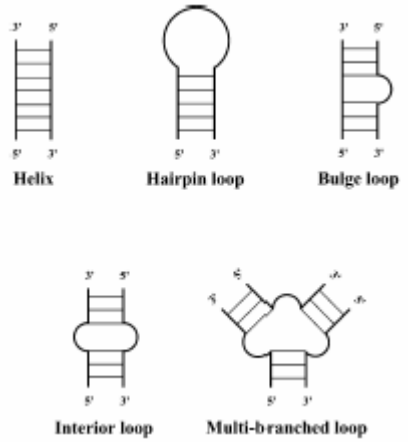

Abbildung 1: *grundlegende Sekundärstrukturelemente in der RNA – Strukturbeispiele und Bezeichnung.*

Allerdings ist die experimentelle Bestimmung der Sekundärstruktur von Nukleinsäuren aufgrund der Flexibilität des Moleküls und seiner Anfälligkeit zu Änderungen der internen Bindungsbeziehungen während der nötigen Präparation extrem schwierig (besonders für langkettige Moleküle), obwohl hier in den letzten Jahren große Fortschritte erzielt wurden. Umso wichtiger ist es also, Alternativen zur Strukturbestimmung zu haben, die diese komplizierte und fehleranfällige Prozedur ersetzen. Hier kann die computergestützte Vorhersage gute Dienste leisten.

Grundidee des Verfahrens

Grundlage des Algorithmus ist die Annahme, dass die Faltung der Nukleinsäure lediglich in die fünf in Abbildung 1 gezeigten einfachen Strukturen erfolgt und in keine weiteren.

Das Verfahren verfolgt den Ansatz der Minimierung der freien Energie der zu faltenden Sequenz. Die freie Energie setzt sich zusammen aus den Beiträgen der inneren Energie des Moleküls sowie der Entropie, unter Einbeziehung der herrschenden absoluten Temperatur. Formel 1 zeigt die Definition der freien Energie.

$$F = U - T * S$$

Formel 1: *Definition der freien Energie F mit U der Energie, T der absoluten Temperatur (Kelvin) und S der Entropie.*

Es ist zwingend nötig, die freie Energie statt lediglich der Energie zu betrachten, da letztere mit steigender Anzahl ausgebildeter Wasserstoffbrückenbindungen zwischen den Basen der Ribonukleotide zunehmend abnimmt. Ideal wäre demnach also ein Zustand, in dem jede Base mit einer anderen eine Wasserstoffbrückenbindung ausbildet. Dieser Zustand ist für RNA aber in keiner experimentell bestimmten Struktur aufgetreten, und er widerspräche auch den Prinzipien der Thermodynamik, nach deren zweiten Hauptsatz die Entropie (als Maß für die „Unordnung“ eines Systems) stets zunimmt, also auch während des Faltungsvorgangs von der Ribonukleotidsequenz mit ungepaarten Basen in die Sekundärstruktur. Bei dem angegebenen Beispiel für die minimale Energie hätte die Ordnung des Systems aber stark zu- und damit die Entropie abgenommen. Es ist also der Realität näher, ein Maß zu betrachten, dass die Energie eines Moleküls mit der Entropie in Beziehung setzt – dies ist die freie Energie.

Allerdings bringt dieser Ansatz einige Probleme mit sich. Einflüsse auf Energie und Entropie, die sich aus möglichen Tertiärstrukturelementen ergeben, bleiben hier, wo nur die Sekundärstruktur betrachtet wird, unberücksichtigt. Weiterhin sind sämtliche Angaben die freie Energie betreffend nur Approximationen – hier könnten leichte Variationen zu anderen Ergebnissen (letztendlich zu anderen Faltungen der RNA) führen.
Darüber hinaus muss es nicht unbedingt sinnvoll sein, zu versuchen, die freie Energie zu minimieren, da die tatsächliche Faltung auch eine unter diesem Aspekt suboptimale Struktur sein kann.

Der Algorithmus

Der Algorithmus arbeitet in zwei Schritten. Für eine gegebene Sequenz sind die verschiedenen Sekundärstrukturen nicht gleich wahrscheinlich. Die Gibbs-Boltzmann-Verteilung einer speziellen Sekundärstruktur I für eine gegebene Sequenz S ist gegeben durch Formel 2. Sie beschreibt die Wahrscheinlichkeit eines Zustandes eines Systems (hier: die Wahrscheinlichkeit der Sekundärstruktur I für die Sequenz S), welches im thermischen Gleichgewicht an die absolute Temperatur T gekoppelt ist. U beschreibt hier die Zustandssumme der Gibbs-Boltzmann-Verteilung, die sich berechnet wie durch Formel 3 beschrieben.

$$P(I) = \frac{\exp\left(\frac{-E(S,I)}{RT}\right)}{U}$$

Formel 2: *die Gibbs-Boltzmann-Verteilung einer Sekundärstruktur I für eine Sequenz S, E(S,I) die freie Energie, R die Gaskonstante, T die absolute Temperatur in Kelvin, U die Zustandssumme.*

$$U = \sum_{I} \exp\left(\frac{-E(S,I)}{RT}\right)$$

Formel 3: *Berechnung der Zustandssumme der Gibbs-Boltzmann-Verteilun., E(S,I) die freie Energie der Sekundärstruktur I für die Sequenz S, R die Gaskonstante, T die absolute Temperatur in Kelvin.*

Im ersten Schritt werden die Zustandssummen aller Teilsequenzen in Abhängigkeit ihrer freien Energie berechnet. Mit ihnen ist es möglich, nach Formel 2 die Gibbs-Boltzmann-Verteilungen der verschiedenen möglichen Sekundärstrukturen einer Teilsequenz zu bestimmen, aus denen im zweiten Schritt rekursiv gesampelt wird.

Schritt 1 – Berechnung der Zustandssummen

Besteht eine Ribonukleinsäure aus n Ribonukleotiden, so sei R_{ij}, $1 = i,j = n$ die Teilsequenz in 5’-3’-Richtung vom Ribonukleotid an Position i bis zu dem an Position j, r_k das Ribonukleotid an Position k, l_{ij} die Sekundärstruktur für R_{ij}, wobei r_i und r_j mit einer Wasserstoffbrücke verbunden sein können oder auch nicht, und IP_{ij} die Sekundärstruktur für R_{ij}, wobei r_i und r_j über eine Wasserstoffbrücke miteinander verbunden sind. Aufsummiert über alle möglichen Sekundärstrukturen des Fragments R_{ij} ergeben sich nach den in Formel 4 gezeigten

Berechnungen zwei Zustandssummen *u(i,j)* und *up(i,j),* wobei *u(i,j)* die Summe über alle möglichen I_{ij}, *up(i,j)* die Summe über alle möglichen IP_{ij} ist. Die Berechnung erfolgt rekursiv durch den Algorithmus von McCaskill.

$$u(i,j) = \sum_{I_{ij}} \exp\left(\frac{-E(R_{ij}, I_{ij})}{RT}\right)$$

$$up(i,j) = \sum_{IP_{ij}} \exp\left(\frac{-E(R_{ij}, IP_{ij})}{RT}\right)$$

Formel 4: *Berechnung der Zustandssummen . u(i,j) summiert über Sekundärstrukturen für R_{ij} mit potentiell ungepaarten terminalen Nukleotiden, up(i,j) über Sekundärstrukturen für R_{ij} mit gepaarten terminalen Nukleotiden. $E(R_{ij},I_{ij})$ bzw. $E(R_{ij},IP_{ij})$ die freie Energie einer Sekundärstruktur für die Teilsequenz R_{ij}, R die Gaskonstante, T = 310,15K.*

Schritt 2 – Sampeln von Sekundärstrukturen aus der Gibbs-Boltzmann-Verteilung

Mit den in Schritt 1 berechneten Zustandssummen lassen sich nun die Sampelwahrscheinlichkeiten für die möglichen Sekundärstrukturen einer (Teil-) Sequenz berechnen. In einer (Teil-) Sequenz R_{ij} kann auf fünf verschiedene Arten eine Wasserstoffbrückenbindung zwischen zwei Basen ausgebildet werden, wie in Abbildung 2 dargestellt.

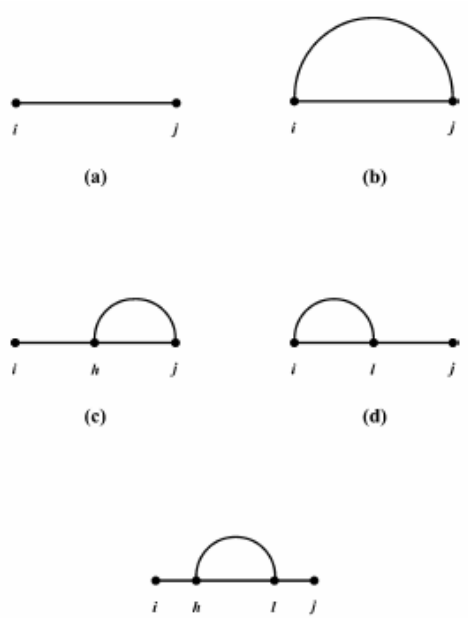

Abbildung 2: *Möglichkeiten zur Ausbildung einer Wasserstoffbrückenbindung in einer Nukleotidsequenz: a) keine Bindung; b) Bindung zwischen den terminalen Nukleotiden; c) Bindung zwischen dem letzten Nukleotid und einem Nukleotid innerhalb der Sequenz; d) Bindung zwischen dem ersten Nukleotid und einem Nukleotid innerhalb der Sequenz; d) Bindung zwischen inneren Nukleotiden.*

Unter der Annahme, dass nicht bekannt ist, ob die terminalen Nukleotide einer Teilsequenz R_{ij} eine Bindung ausbilden oder nicht, ergeben sich für diese fünf Möglichkeiten unter Nutzung der in Schritt 1 berechneten Zustandssummen *u(i,j)* und *up(i,j)* die Sampelwahrscheinlichkeiten

$$P_0 = 1/u(i,j)$$

$$P_{ij} = up(i,j)\exp[-etp(i,j)/RT]/u(i,j)$$

$$P_{hj} = up(h,j)\exp\{-[ed5(h,j,h-1) + etp(h,j)]/RT\}/u(i,j),\ i < h < j,$$

$$P_{il} = up(i,l)\exp[-etp(i,l)/RT]\{\exp[-ed3(i,l,l+1)/RT]u(l+2,j) + u(l+1,j) - u(l+2,j)\}/u(i,j),\ i < l < j$$

$$P_{s1h} = s1(h,j)/u(i,j),\ i < h < j-1$$

$$P_{hl} = up(h,l)\exp\{-[ed5(h,l,h-1) + etp(h,l)]/RT\}\{\exp[-ed3(h,l,l+1)/RT]u(l+2,j) + u(l+1,j) - u(l+2,j)\}/s1(h,j),\ h < l < j$$

wobei P_0 die Sampelwahrscheinlichkeit für den Fall a) ist, P_{ij} die für Fall b), $\{P_{hj}\}$ die für Fall c), $\{P_{il}\}$ die für Fall d) $\{P_{s1h}\}$ die Wahrscheinlichkeiten, Nukleotid r_h in Fall e) zu sampeln und $\{P_{hl}\}$ die Wahrscheinlichkeiten, Nukleotid r_l in Fall e) zu sampeln, nachdem r_h festgelegt wurde, wobei $i < h < l < j$ gelten muss. Zu beachten ist, dass es sich in den letzten vier Fällen jeweils um eine Menge von Wahrscheinlichkeiten handelt, da es jeweils mehr als eine Möglichkeit gibt, diese Bindungsbeziehung auszubilden. *etp()* beschreibt einen Strafterm für unzulässige Basenpaarungen, *ed5()* beschreibt die freie Energie eines dangling 5'-Endes, analog dazu *ed3()* die eines dangling 3'-Endes.
Bilden die Nukleotide r_i und r_j eine Wasserstoffbrückenbindung aus, so kann diese Bestandteil einer der fünf Sekundärstrukturen sein, die in Abbildung 1 vorgestellt wurden. Die Sampelwahrscheinlichkeiten berechnen sich unter dieser Annahme wie folgt:

$$Q_{ijH} = \exp[-eh(i,j)/RT]/up(i,j)$$

$$Q_{ijS} = \exp[-es(i,j,i+1,j-1)/RT]up(i+1,j-1)/up(i,j)$$

$$Q_{ijBI} = \{\Sigma_{i<h<l<j}\exp[-ebi(i,j,h,l)/RT]up(h,l)\}/up(i,j)$$

$$Q_{ijM} = up_m(i,j)/up(i,j)$$

$$Q_{hlBI} = \exp[-ebi(i,j,h,l)/RT]up(h,l)/\{\Sigma_{i<h'<l'<j}\exp[-ebi(i,j,h',l')/RT]up(h',l')\},\ i < h' < l' < j$$

Dabei ist Q_{ijH} die Sampelwahrscheinlichkeit für einen Hairpin Loop, Q_{ijS} die für ein stacking Basepair, Q_{ijBI} die für einen Bulge oder Interior Loop und Q_{ijM} die Sampelwahrscheinlichkeit für einen Multibranched Loop. Die $\{Q_{hlBI}\}$ werden genutzt, die betroffenen Nukleotide r_h und r_l zu sampeln, nachdem durch den Sampelprozess ein Bulge oder Interior Loop als Sekundärstruktur ausgewählt wurde. *eh()* beschreibt dabei die freie Energie eines Hairpin, *es()* die eines stacking Basepairs und *ebi()* die eines Bulge bzw. Interior Loops.

Der Sampelprozess läuft nun ab wie folgt: zunächst werden für eine (Teil-) Sequenz R_{ij} in Abhängigkeit davon, ob das Bindungsverhältnis zwischen r_i und r_j bekannt ist oder nicht die entsprechenden Wahrscheinlichkeiten P_x oder Q_x unter Nutzung der in Schritt 1 berechneten Zustandssummen *u(i,j)* bzw. *up(i,j)* berechnet. Dann wird gemäß diesen Wahrscheinlichkeiten eine Möglichkeit aus den Bindungsbeziehungen (es war unbekannt, ob r_i und r_j paaren, zuvor die P_x berechnet) bzw. der Sekundärstrukturen (es war bekannt, dass r_i und r_j paaren, zuvor Q_x berechnet) ausgewählt. Der Algorithmus nutzt dazu zwei Stacks, die nachfolgend mit A und B bezeichnet werden. Auf Stack A werden Tupel der Form (i,j,l) abgelegt, die ein noch zu faltendes Teilstück, genauer die Teilsequenz R_{ij} repräsentieren. I symbolisiert die Kenntnis über die Bindungsbeziehung zwischen r_i und r_j und kann den Wert 0 bzw. 1 annehmen. I=0 symbolisiert, dass unbekannt ist, ob r_i und r_j paaren, I=1 steht dafür, dass r_i und r_j miteinander über eine Wasserstoffbrückenbindung verbunden sind. Stack B wird nach und nach durch den Algorithmus aufgefüllt und enthält Informationen über den Bindungszustand der Nukleotide. So enthält nach Terminierung des Verfahrens Stack B alle Angaben darüber, welches Nukleotid ungepaart bzw. gepaart ist, und mit welchem anderen Nukleotid diese Bindung eingegangen wird. Dies ist eine Repräsentation der Sekundärstruktur der zu faltenden RNA. Der Algorithmus startet mit dem Tupel (1,n,0) auf Stack A. Es ist nicht bekannt, ob r_1 und r_n miteinander paaren (I=0), also werden die P_x als Sampelwahrscheinlichkeiten gebildet und eine der fünf Möglichkeiten ausgewählt. Für den Fall, dass aus Abbildung 2 Fall a) ausgewählt wurde, werden alle Nukleotide r_1 bis r_n mit der Information „ungepaart“ auf Stack B eingefügt und der Algorithmus terminiert. Für Fall b) wird das Tupel (1,n,1) auf Stack A abgelegt, da die gesamte Sequenz R_{1n} immer noch zu falten ist, nun aber bestimmt wurde, dass r_1 und r_n über eine Wasserstoffbrückenbindung miteinander verbunden sein sollen. Für Fall c) wird das Tupel (h,n,1) auf Stack A abgelegt, da dieses Teilstück auch weiterhin noch zu falten ist und r_h und r_n miteinander paaren sollen; gleichzeitig werden die Nukleotide r_1 bis r_{h-1} mit der Information „ungepaart“ aus Stack B abgelegt. Wurde Fall d) ausgewählt, werden die Tupel (1,l,1) und (l+1,n,0) auf Stack A eingefügt. Für Fall e) werden (h,l,1) und (l+1,n,0) auf Stack A eingefügt sowie die Nukleotide r_1 bis r_{h-1} als ungepaart auf Stack B abgelegt. Im nächsten Schritt wird nun das nächste Tupel von Stack A genommen. Sollte wieder I=0 gelten, wird erneut verfahren wie zuvor beschrieben. Ist aber I=1, so werden die Q_x als

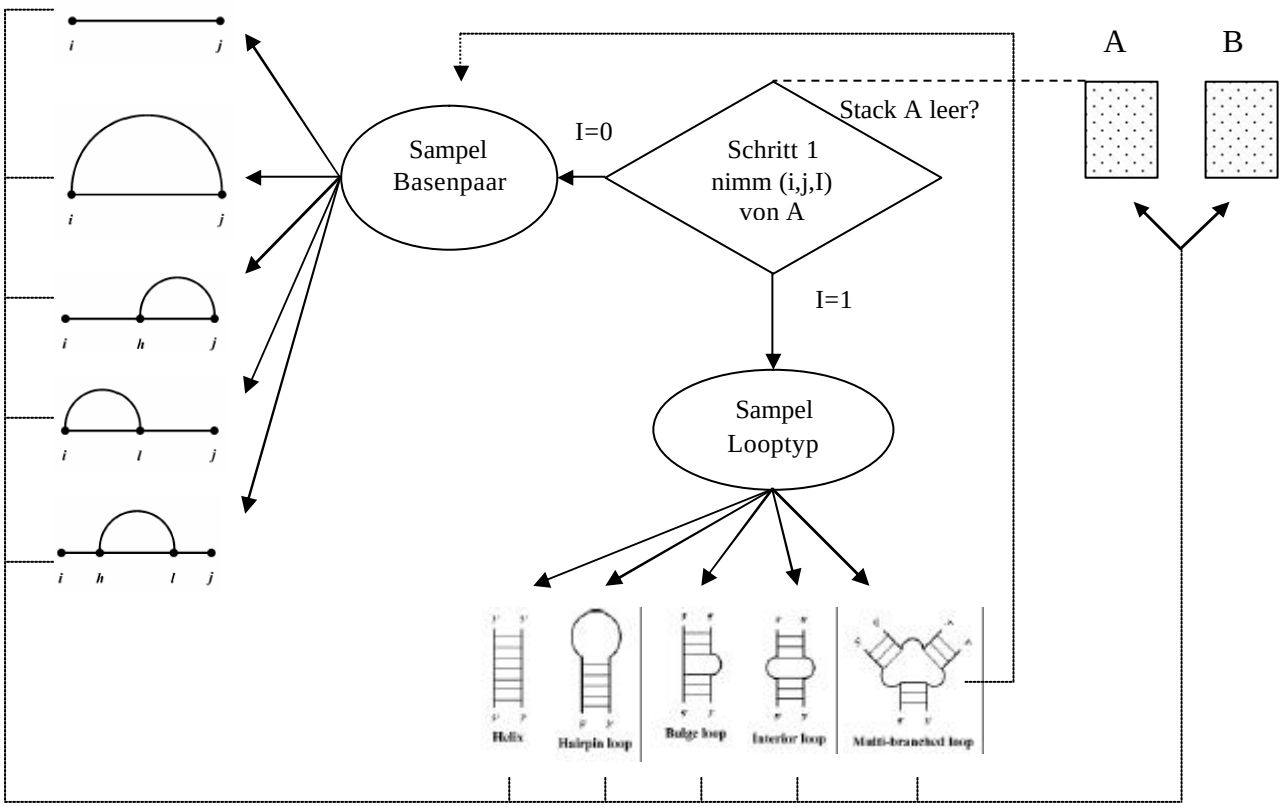

Abbildung 3: *Sampling im Überblick. Durch das Sampeln des Basenpaars wird eine der Möglichkeiten zur Paarung ausgewählt, durch das Sampeln des Looptyps wird bestimmt, zu welchem Sekundärstrukturelement ein Basenpaar zugehörig ist. Im Fall des Multibranched Loop wird für jedes interne Basenpaar die Bindungsmöglichkeit gesampelt. Das Verfahren terminiert, sobald Stack A leer ist.*

Sampelwahrscheinlichkeiten gebildet und ihnen entsprechend bestimmt, welche Sekundärstruktur diesem Teilstück zugeordnet wird. Für den Fall eines Hairpins werden die gepaarten bzw. ungepaarten Nukleotide auf Stack B abgelegt. Für eine Helix wird das Basenpaar r_i-r_j auf Stack B angelegt sowie das Tupel (i+1,j-1,1) in Stack A eingefügt. Für einen Bulge bzw. Interior Loop werden zunächst die $\{Q_{hlBl}\}$ berechnet und die Nukleotide r_h und r_l gesampelt. Dann wird das Basenpaar r_i-r_j sowie die ungepaarten Basen des Loops in Stack B eingefügt und das Tupel (h,l,1) auf Stack A eingefügt. Im Fall des Multibranched Loop wird nun das erste innere Basenpaar nach Bildung der P_x gesampelt, dann das zweite und so fortgefahren, bis alle inneren Basenpaare abgearbeitet sind. Nach jedem der zuvor beschriebenen Fälle wird das nächste Tupel von Stack A genommen. Das Verfahren terminiert, wenn Stack A leer ist. Abbildung 3 zeigt den Vorgang des Sampling im Überblick.

Die Wahrscheinlichkeit einer Sekundärstruktur nimmt mit wachsender freier Energie exponentiell ab, bedingt durch die Gibbs-Boltzmann-Verteilung. So wird während des Samplings mit hoher Wahrscheinlichkeit die Sekundärstruktur mit minimaler freier Energie erzeugt, mit relativ hoher Wahrscheinlichkeit eine Sekundärstruktur, deren freie Energie nahe der der Sekundärstruktur mit minimaler freier Energie liegt, und mit niedriger Wahrscheinlichkeit eine Sekundärstruktur mit einer hohen freien Energie. Dies ist ein entscheidender Vorteil im Vergleich zu Verfahren, die lediglich die Sekundärstruktur mit minimaler freier Energie bilden, da die in dieser Hinsicht optimale Faltung

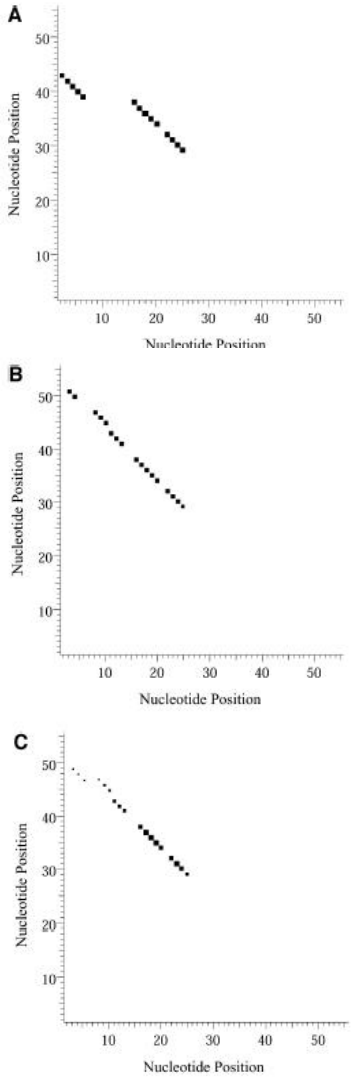

Abbildung 4: *2D-Histogramme für die Klassen 1A, 1B und 1C. Die Achsen geben die Nukleotidpositionen an, die Quadrate symbolisieren des vorhandensein von Bindungen, die Intensität der Quadrate die Häufigkeit des Auftretens dieser Bindung.*

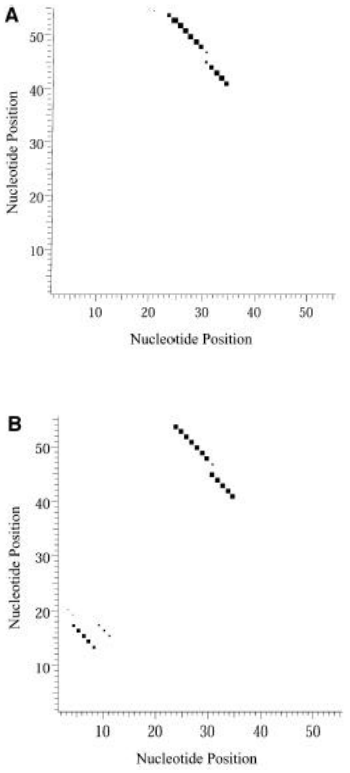

Abbildung 5: *2D-Histgoramme für Die Klassen 2A und 2B. Die Achsen geben die Nukleotidpositionen an, die Quadrate symbolisieren des vorhandensein von Bindungen, die Intensität der Quadrate die Häufigkeit des Auftretens dieser Bindung.*

generell nur selten die native Faltung eines erzeugt aber wie oben beschrieben mit relativ hoher Wahrscheinlichkeit auch solche suboptimalen Faltungen, kann also auch die reale Faltung generieren, wenn diese nicht die mit minimaler freier Energie ist.

Anwendungsbeispiele

Nachfolgend soll an zwei Beispielen das Potential des Algorithmus demonstriert werden, indem zwei RNAs mit bereits bekannter Struktur von ihm gefaltet und die generierten Ergebnisse bewertet werden.

Leptomonas collosoma

Der Flagellat *Leptomonas collosoma* verfügt über eine 56 Nukleotide lange „spliced leader" RNA (SL RNA), von der bekannt ist, dass sie *in vivo* in zwei Faltungen bekannter Struktur vorkommt. Nach Berechnung der Zustandssummen wurde 1000 mal gesampelt und die entstandenen Sekundärstrukturen miteinander verglichen. Die entstandenen Sekundärstrukturen lassen sich nach ihren Strukturmerkmalen zwei Klassen zuordnen (1 und 2), wobei Klasse 1 noch in drei weitere Unterklassen (A, B und C), Klasse 2 in zwei weitere Unterklassen (A und B) unterteilt werden kann.

Die Strukturen in Klasse 1 (Abbildung 4) verfügen über zwei identische Helices ($U^{16}A^{38}$, $G^{17}C^{37}$, $U^{18}A^{36}$, $A^{19}U^{35}$, $G^{20}C^{34}$ und $U^{22}A^{32}$, $C^{23}G^{31}$, $A^{24}U^{30}$, $G^{25}U^{29}$). Den Unterklassen B und C sind 2 weitere kleine Helices gemein, sie unterscheiden sich jedoch im Vorhandensein eines kurzen Hairpins. Unterklasse B repräsentiert die Struktur, die von Moleküls ist. Der hier vorgestellte Algorithmus mfold(3.1) als die Struktur mit minimaler freier Energie erzeugt wurde.
Auch die Strukturen der Klasse 2 (Abbildung 5) besitzen jeweils zwei identische Helices. Die Unterklassen A und B unterscheiden sich durch einen zusätzlich vorhandenen Stem am 5'-Ende der Unterklasse B.

Um die Unterschiede und Gemeinsamkeiten der verschiedenen Klassen besser betrachten zu können, beschränken sich die weitergehenden Analysen auf die Repräsentanten der einzelnen Klassen, also die Strukturen, die in jeder Klasse mit der größten absoluten Häufigkeit erzeugt wurden (Abbildung 6). Drei der fünf Repräsentanten sind besonders interessant: der der Klasse 1A ist die eine der beiden bekannten nativen Faltungen der SL RNA von *L. collosoma*; der Repräsentant der Klasse 1B stellt die von mfold(3.1) generierte Faltung mit der minimalen freien Energie dar, und der Repräsentant der Klasse 2A ist die zweite bekannte native Faltung der SL RNA. Der Repräsentant der Klasse 1C ist bis auf eine kurze Helix mit der mfold-Struktur identisch.

Abbildung 7 zeigt eine Übersicht über die Häufigkeit der Klassen sowie deren Repräsentanten. Zu beachten ist, dass die beiden bekannten Faltungen der untersuchten SL RNA (1A und 2A) nur zu sehr geringen Prozentsätzen erzeugt wurden. Klasse 1C, die eine leichte Variation der Struktur mit minimaler Energie (1B) darstellt, wurde mit Abstand am häufigsten erzeugt, gefolgt von der mfold-Struktur (1C). Dies zeigt, dass der Algorithmus wie gewünscht Strukturen generiert, deren freie Energie

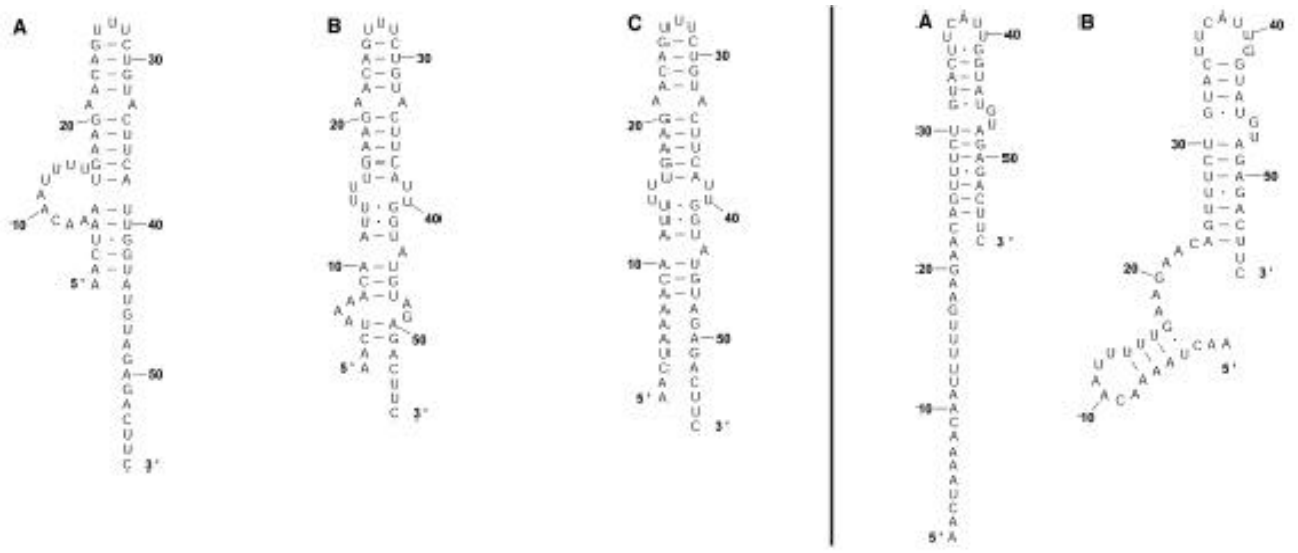

Abbildung 6: *Repräsentanten der Klassen 1A, 1B, 1C (links) und 2A, 2B (rechts).1A: erste native Faltung; 1B: Faltung mit minimaler freier Energie; 2A: zweite native Faltung.*

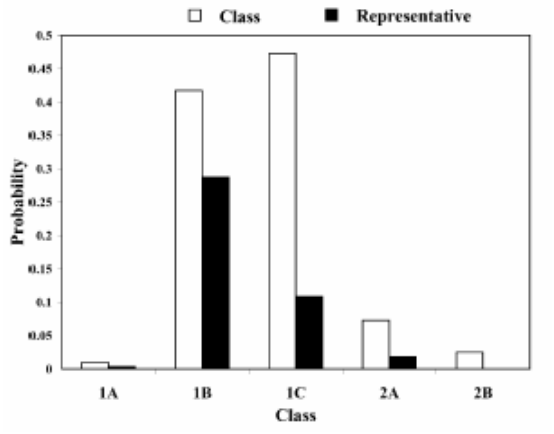

Abbildung 7: *Das Diagramm zeigt die relativen Häufigkeiten der einzelnen Klassen und deren Repräsentanten.*

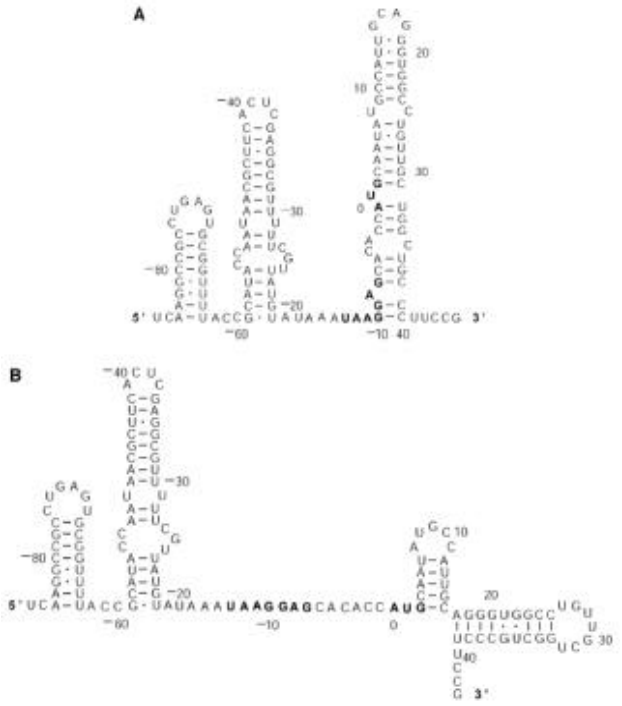

Abbildung 8: *die beiden Konformationen der cIII-mRNA des Bakteriophagen l. A: keine Translation möglich. B: Translation möglich.*

geringfügig über der minimalen freien Energie liegt. Es wird aber auch deutlich, dass das Erzeugen suboptimaler Lösungen nicht zwangsläufig zu den realen Strukturen führen muss, denn die beiden tatsächlichen Faltungen der SL RNA wurden nur in sehr geringem Maß gebildet. Allerdings ist der Vorteil dieser Methode gegenüber der Berechnung der Struktur mit der minimalen freien Energie ebenfalls klar ersichtlich, denn mit diesem Ansatz wäre es nie zur Generierung der beiden realen Faltungen gekommen.

Bakteriophage l

Im zweiten Beispiel soll die Struktur der mRNA des cII-Gens des Bakteriophagen λ vorhergesagt werden. Sie verfügt über zwei charakteristische Bereiche: das Startcodon an den Positionen 0 bis 3 und die Shine-Dalgarno-Sequenz an der Position -13 bis -7, die im Andockungs-Vorgang des Ribosoms an die mRNA involviert ist. Offensichtlich ist, dass sowohl das Startcodon als auch die Shine-Dalgarno-Sequenz zugänglich sein müssen, damit die mRNA translatiert werden kann – dies bedeutet, dass sie nicht in Sekundärstrukturelemente eingebunden sein dürfen. Wie die SL RNA von Leptomonas collosoma kommt auch die cIII-mRNA des Bakteriophagen λ in zwei Konformationen vor (Abbildung 8). In der ersten Konformation sind sowohl das Startcodon als auch die Shine-Dalgarno-Sequenz teilweise in Sekundärstrukturelemente eingebunden, folglich kann in dieser Faltung die Translation nicht initiiert werden. In der zweiten Konformation sind die beiden Sequenzbereiche hingegen gänzlich ungepaart, was das Andocken des Ribosoms und den Beginn der Translation ermöglicht. Der Sampling-Schritt des Algorithmus wurde nun 100 mal wiederholt und die erzeugten Faltungen von Hand betrachtet und charakterisiert. 89 der 100 Strukturen waren leichte Variationen der Struktur A. Von diesen 89 war in 67 der linke der Stems enthalten, in 72 wurde der rechte Stem nahezu exakt vorhergesagt; gelegentlich wurde er um zwei Basenpaare verlängert. 3 der 100 Strukturen waren Varianten der Struktur B, wobei die Shine-Dalgarno-Sequenz eine zwei Basenpaare lange zusätzliche Helix enthielt. Die 8 übrigen Strukturen zeigten mit keiner der bekannten Konformationen eine ausreichende Übereinstimmung. Insgesamt waren also 92 der 100 erzeugten Faltungen zu einem Großteil übereinstimmend mit den bekannten Faltungen.

Als Ergebnis dieser beiden Fallbeispiele kann gesagt werden, dass der Algorithmus wie beabsichtigt Faltungen erzeugt, die in ihrer freien Energie minimal von der der in dieser Hinsicht optimalen Faltung abweicht.
Zu beachten bleibt aber, dass nicht jede suboptimale Faltung die tatsächliche (suboptimale) Faltung der RNA ist. Die Autoren versprechen sich bessere Ergebnisse, wenn die Tertiärstruktur in die Betrachtung miteinbezogen würde. Allgemein scheint der Algorithmus besser zur Faltungsvorhersage für mRNA als für funktionelle RNA geeignet zu sein.

Weitere Anwendungen

Probability Profiling

Einzelsträngige (also ungepaarte) RNA-Regionen können mit anderen Nukleinsäuren und auch mit Proteinen interagieren. Mit Hilfe des Sampling-Schritts des von Ding und Lawrence Algorithmus ist es möglich, solche „accessible sites“ in RNA-Molekülen vorherzusagen und zu bestimmen. Dazu wird ein „Probability Profile“ erstellt, ein Diagramm, in dem die Wahrscheinlichkeit, dass die nächsten w Nukleotide nicht gepaart sind, gegen die Sequenzpositionen aufgetragen wird. Dies ist möglich, indem der Sampling-Schritt des Algorithmus mehrfach durchgeführt wird. Analysiert man, in wieviel Prozent der Fälle das Nukleotid an Position i nicht gepaart war und multipliziert diese Häufigkeit mit den entsprechenden Werten der nachfolgenden w-1 Nukleotide, erhält man den im Probability Profile an Position i einzutragenden Wert. Der sogenannte „ss-count“, der nachfolgend als Vergleich verwendet wird, betrachtet lediglich die Häufigkeit der Paarung eines einzelnen Nukleotids, ohne die der nachfolgenden miteinzubeziehen. Abbildung 9 zeigt das Probability Profile für die γ-Glutamyl-Hydrolase von *Homo sapiens.*

Die von mfold gelieferte Struktur mit der minimalen freien Energie liefert keine Hinweise zur Vorhersage von accessible sites, da sie eine feststehende Struktur ist, mit der nur eine binäre Aussage (Position i gepaart oder ungepaart) getroffen werden kann. Der ss-count betrachtet wie bereits beschrieben die Statistik jedes Nukleotids isoliert, ohne die Information über die umgebenden Sequenzabschnitte zu berücksichtigen. Das kann zu irreführenden Angaben hinsichtlich der Zugänglichkeit eines Bereichs der Nukleinsäure führen, da die isolierte Betrachtung nachfolgender Nukleotide höhere Wahrscheinlichkeiten für die Zugänglichkeit eines Bereiches liefert als die Betrachtung eines Nukleotids unter der Bedingung, dass auch die darauffolgenden ungebunden sind. Daher ist die Aussage des Probability-Profiles die verlässlichste der drei hier kurz vorgestellten Varianten, auch wenn die Intensität der Vorhersage bei dieser Methode am geringsten ist.

Bisher wurde nur die Information berücksichtigt, ob ein Nukleotid mit einem anderen über eine Wasserstoffbrückenbindung verbunden ist oder nicht. Während des Sampling-Schritts des Algorithmus wird allerdings noch wesentlich mehr Information generiert: es wird festgelegt, in welchem Looptyp, die entsprechenden (gepaarten oder ungepaarten) Nukleotide enthalten sind. Damit ist das Probability Profiling für die einzelnen Looptypen getrennt möglich. Die entstehenden Loop-Probability-Profiles geben für eine festzulegende Schrittweite w die Wahrscheinlichkeit an, dass ab Position i w Nukleotide in einem festzulegenden Looptyp ungepaart sind. Abbildung 10 zeigt die Probability Profiles für Haipin Loop, Multibranched Loop sowie Exterior Loop für die Alanin-tRNA von *Escherichia coli*.

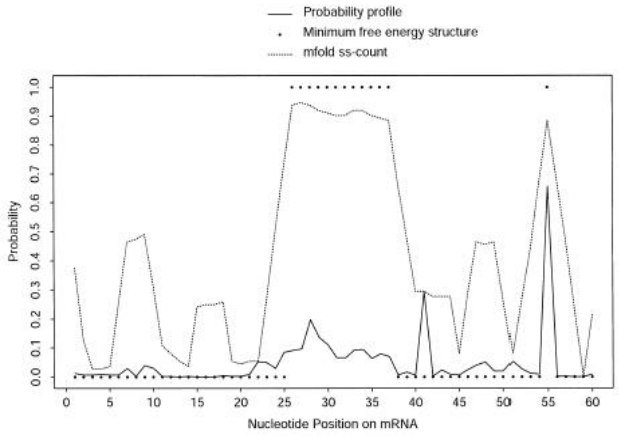

Abbildung 9: *Probability Profile für die Homo sapiens g-Glutamyl-Hydrolase. Gesampelt wurde 1000 mal, w=4.*

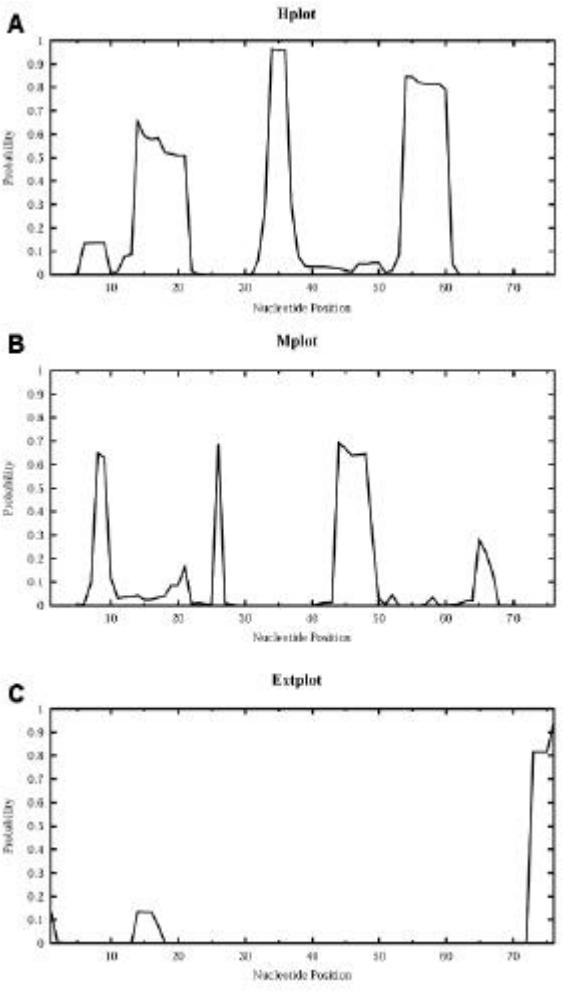

Abbildung 10: *Loop-Probability-Profiles für Hairpin Loop (A), Multibranched Loop (B) und Exterior Loop (C) der Alanin-tRNA von Escherichia coli.*

Die für tRNA charakteristischen Loop-Strukturen sind deutlich zu erkennen. So zeigt Abbildung 10 A deutlich erkennbar den TΨC-Loop, den Anticodon-Loop sowie den D-Loop. In Abbildung 10 B sind die ungepaarten Bereiche des zentralen Multibranched Loop, der durch die tRNA gebildet wird, klar zu erkennen. Abbildung 10 C hingegen zeigt mit einem deutlichen Peak das freie aus 3 Nukleotiden bestehende 3'-Ende der tRNA.

Accessibility Plots

Damit verschiedene Nukleinsäuren miteinander interagieren können, müssen sie über ungepaarte Bereiche verfügen. Im Idealfall sind diese gleich lang. Die durch Probability Profiles gegebenen Informationen geben jeweils nur Vorhersagen für den Zustand eines einzelnen Moleküls. Um mögliche Interaktionen vorherzusagen, können die Probability Profiles von Molekülen überlagert werden. Abbildung 11 zeigt einen so entstandenen „Accessibility Plot" für die *Homo sapiens* γ-Glutamyl-Hydrolase-mRNA und ein künstlich erzeugtes Antisense-Molekül.

Da sich die Moleküle gegeneinander verschieben können oder die die einzelsträngigen Bereiche umgebenden RNA-Domänen gänzlich anders gefaltet und unterschiedlich lang sein können, müssen die potentiell zur Interaktion befähigten Bereiche nicht an den selben Sequenzpositionen innerhalb der einzelnen Moleküle liegen. Es ist jedoch darauf zu achten, dass sie eine maximale Sequenzkomplementarität aufweisen, da es sonst zu keiner Interaktion kommen kann.

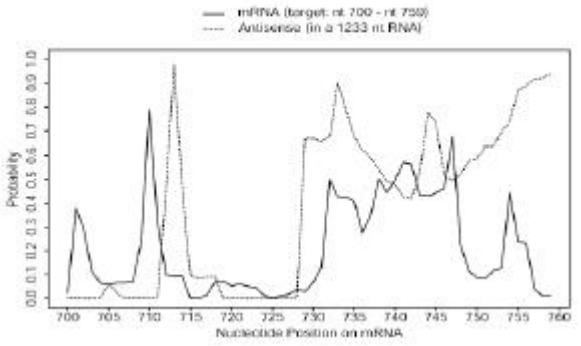

Abbildung 11: *Accessibility Plot für die Homo sapiens g-Glutamyl-Hydrolase-mRNA und ein künstlich erzeugtes 1233 Nukleotide langes Antisense-Molekül. Für die einzelnen Probability Profiles wurde w=4 gewählt, es wurde 1000 gesampelt. Das Antisense-Stück ist 60 Nukleotide lang und in ein insgesamt 1233 Nukleotide langes RNA-Molekül eingebettet. Die Probability Profiles wurden gegeneinander verschoben, bis es zur Übereinstimmung der ungepaarten Bereiche kam.*

Repräsentative Sampelgrößen

In den aufgeführten Beispielen wurde der Sampling-Schritt meist 1000 mal wiederholt. Doch ist das eine Samplegröße, mit der repräsentative Aussagen getroffen werden können? Um dies herauszufinden wandten sich Ding und Lawrence erneut der *Homo sapiens* γ-Glutamyl-Hydrolase-mRNA zu. Sie hat eine länge von 1187 Nukleotiden, wodurch es ca. 10^{303} mögliche Sekundärstrukturen gibt – zum Vergleich: die Anzahl der Atome im Universum wird auf 10^{78} geschätzt. Es wurden nun 2 Sammlungen von Sekundärstrukturen erstellt. Für jede Sammlung wurde der Samplingschritt 1000 mal wiederholt. Anschließend wurde für jede der Sammlungen ein 2D-Histogramm erstellt, in dem die Bindungsbeziehungen zwischen den Nukleotiden der Sammlung eingetragen wurden. Abbildung 12 zeigt die Überlagerung der beiden Histogramme. Sie sind offensichtlich nahezu identisch. Also können bei einer Samplegröße von 1000 statistisch repräsentative Aussagen getroffen werden.

Fazit

Ye Ding und Charles Lawrence haben in ihrer Veröffentlichung „A statistical sampling algorithm for RNA secondary structure prediciton" einen Algorithmus vorgestellt, der in einer Laufzeit von $O(n^3)$ und einem Speicherbedarf von $O(n^2)$ in der Lage ist, Sekundärstrukturen einer gegebenen RNA-Sequenz ohne Berücksichtigung von Pseudoknoten vorherzusagen, deren freie Energie

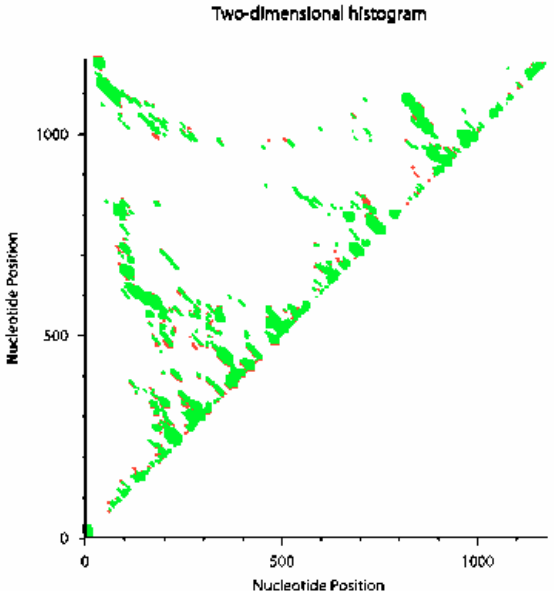

Abbildung 12: *Überlagerung der eingefärbten 2D-Histogramme von zwei Strukursammlungen, die jeweils 1000 vorhergesagte Sekundärstrukturen für die Homo sapiens g-Glutamyl-Hydrolase-mRNA enthalten.*

nahe der der Sekundärstruktur mit minimaler freien Energie liegt. Wie sinnvoll allerdings ein Ansatz ist, der Sekundärstrukturen ohne Pseudoknoten erzeugt, obwohl diese als Strukturmerkmale in RNA-Molekülen vorkommen, müssen weitere Untersuchungen zeigen. Das Programm steht im Paket sfold zur freien Verfügung.

„Noncoding RNA gene detection using comparative sequence analysis"
(Elena Rivas, Sean Eddy)

Motivation

Neben der mRNA, die nur ein während der Genexpression entstehendes Intermediat ist und letztlich zerfällt bzw. abgebaut wird, gibt es noch weitere Formen der RNA, die aufgrund ihrer Sekundärstruktur in biologischen Prozessen eine wichtige Rolle spielen. Sie tragen keine Information für Proteinsequenzen, sondern haben wie zum Beispiel Ribozyme katalytische Funktion oder sind an anderen Vorgängen in der Zelle beteiligt, wie die tRNA an der Translation. Da auch diese funktionelle, nicht für Proteine kodierende RNA (ncRNA) auf der DNA kodiert ist, bedeutet das weiterhin, dass nicht alle Gene auf der DNA für Proteine kodieren, sondern einige für ncRNA. Elena Rivas und Sean Eddy wandten sich dem bisher kaum beachteten Problem zu, solche ncRNA-Gene in DNA-Sequenzen zu erkennen.

Ansätze

Bereits 1988 widmeten sich Maizel und seine Mitarbeiter diesem Problem. Ihr Idee war, dass die Funktion eines Moleküls, also auch die von RNA, durch die Sekundärstruktur bestimmt wird. Damit ncRNA ihre Funktion ausführen kann, darf also die Sekundärstruktur nicht verändert werden (was zu einem Funktionsverlust führen würde), muss also sehr stabil sein. Eine äusserst stabile Sekundärstruktur könnte eine Sekundärstruktur sein, deren freie Energie sehr nahe an der minimalen freien Energie liegt. Demnach müsste die Faltungsvorhersage für solche RNAs wenig divers sein, da ein Ergebnis mit einer großen Anzahl von Strukturen bedeuten würde, dass viele verschiedene Faltungen möglich wären, von denen aber einige eine höhere freie Energie hätten als andere. Man hätte eine ncRNA gefunden, wenn für eine Sequenz bedeutend weniger verschiedene Sekundärstrukturen vorhergesagt werden als für zufällig aufgebaute Sequenzen mit den gleichen Nukleotidhäufigkeiten. Die Ergebnisse dieses Ansatzes waren enttäuschend.

Badger und Olsen entwickelten 1999 für ein anderes Problem einen Lösungsansatz, der auch für die Untersuchung von DNA auf ncRNA-Gene vielversprechend erscheint. Sie versuchten, in Bakteriengenomen (für Proteine) kodierende Bereiche zu erkennen. Dazu arbeiteten sie mit Alignments von Genomen (eng) verwandter Bakterienarten. Sie nutzten blastn, um die Sequenzidentität zwischen den beiden Spezies zu bestimmen, und ihr „CRITICA" getauftes Programm analysierte die Mutationsmuster der gaplos alignierten Bereiche, die die Evolution von einem Genom zum anderen erklärten. Synonyme Mutationen (Mutationen, die nicht die in einem Codon codierte Aminosäure verändern – möglich aufgrund der Redundanz des genetischen Codes) wurden mit einem positiven Score bewertet, nicht synonyme Mutationen mit einem negativen Score. Je höher nun der Gesamtscore war, als desto näher wurde die Verwandtschaftsbeziehung zwischen den beiden Arten eingestuft.
Das von Rivas und Eddy entwickelte Verfahren baut auf dem Ansatz von Badger und Olsen auf, nur dass nun nicht mehr für Proteine kodierende Bereiche, sondern für ncRNA kodierende Bereiche identifiziert werden sollen. Ziel ist es, Bereiche innerhalb eines Alignments von Sequenzen (mit Gaps) den Kategorien „kodierend für Proteine", „kodierend für ncRNA" oder „nicht kodierend" zuzuordnen. Dazu werden Muster aus verschiedenen Mutationsmöglichkeiten betrachtet, die den Übergang der einen Sequenz in die andere erklären können. Die möglichen Arten der Mutation sind: synonyme Mutation (ein Codon kodiert trotz Mutation für die selbe Aminosäure), kompensierende Mutation (Sekundärstruktur wird nicht verändert) und zufällige Mutation. Abbildung 13 zeigt Beispiele für die drei Mutationstypen. Für jeden dieser drei Mutationstypen wird ein Modell gebaut, das in der Lage ist, entsprechend dieses Mutationstyps miteinander verwandte (DNA-) Sequenzen (bzw. Alignments) zu erzeugen. Das Modell von Badger und Olsen wird dahingehend erweitert, als dass es möglich ist, Teile der Alignments

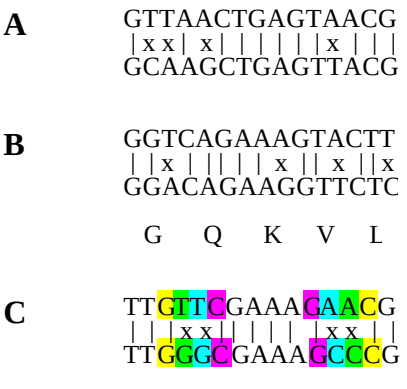

Abbildung 13: *Beispiele für die verschiedenen Sequenztypen. A: Evolution durch positionsunabhängige Mutationen. B: Evolution durch synonyme Mutationen – die codierte Aminosäure pro Codon bleibt identsich. C: komepnsierende Mutation – zwischen den farblich gekennzeichneten Positionen innerhalb einer Sequenz ist jeweils die Ausbildung einer Wasserstoffbrückenbindung möglich.*

zu klassifizieren, die von nicht kodierenden Flanken umgeben sind. Mittels bereits bekannter Algorithmen (Viterbi, Forward, CYK, Inside) wird dann für das eingegebende Alignment und jedes der drei Modelle die Wahrscheinlichkeit ermittelt, dass das Alignment von dem betreffenden Modell generiert wurde, also dem durch dieses Modell repräsentierten Sequenztyp (für Proteine kodierend, für ncRNA kodierend oder für nichts kodierend) zuzuordnen ist. Dies geschieht durch die Berechnung der Bayes'schen Posteriori-Wahrscheinlichkeit für das Alignment und jedes Modell.
Mittels dieses Verfahrens sollen RNAs mit stark konservierter Sekundärstruktur gefunden werden. Da allerdings auch manche mRNAs zum Beispiel zur Translationskontrolle über solche Elemente verfügen, können letztendlich nur ncRNA-Gen-„Kandidaten" gefunden werden, die weiter auf ihre Funktion überprüft werden müssen.

Die Modelle

IID

Das IID Modell ist keines der bereits genannten Modelle. Es dient lediglich dazu, unabhängige Flanken um die Bereiche der durch die drei Modelle zu erzeugen, bzw. um diese durch eingefügt unverwandte Bereiche zu trennen. Es handelt sich um ein einfaches pair-Hidden-Markov-Modell mit 8 Emissionswahrscheinlichkeiten vom Typ $p_s(a)$ mit $s \in \{X, Y\}$, die die Wahrscheinlichkeit angeben, dass Nukleotid a in Sequenz s emittiert wird.

OTH

Das OTH-Modell repräsentiert die Evolution durch zufällige, von anderen Positionen unabhängige, Mutationen. Damit repräsentiert es den Fall, dass das analysierte bzw. vom Modell generierte Alignment weder für Proteine noch für ncRNA kodiert. Es ist ein pair-Hidden-Markov-Model und verfügt über 16 Emissionswahrscheinlichkeiten $p_{OTH}(a,b)$, die jeweils die Wahrscheinlichkeit angeben, dass Nukleotid a in Sequenz X, Nukleotid b in Sequenz Y emittiert wird. Die Gesamtwahrscheinlichkeit des Alignments ergibt sich folglich aus der simplen Multiplikation der Einzelwahrscheinlichkeiten der gegeneinander alignierten Positionen. Abbildung 13 zeigt ein Piktogramm zu diesem Modell.

COD

Das COD-Modell repräsentiert die Evolution durch synonyme Mutationen, erzeugt also alignierte Sequenzen, die für (die selben) Proteine kodieren. Es handelt sich dabei um ein pair-Hidden-Markov-Modell mit 64*64 Emissionswahrscheinlichkeiten $p_{COD}(a_1a_2a_3, b_1b_2b_3)$, die jeweils die Wahrscheinlichkeit angeben, dass das Codon aus den Nukleotiden a_1, a_2 und a_3 in Sequenz X und das Codon bestehend aus den Nukleotiden b_1, b_2 und b_3 in Sequenz Y emittiert wird. Die Wahrscheinlichkeit des Alignments für einen festen Reading-Frame ist dann das Produkt aus den Doppel-Codon-Wahrscheinlichkeiten. Der Reading-Frame ist jedoch nicht bekannt, so muss über alle sechs möglichen Reading-Frames summiert werden. Die Berechnung der Alignment-Wahrscheinlichkeit zeigt Formel 5.

$$P(\overline{XY} \mid COD) = \sum_f P(\overline{XY} \mid f, COD) * P(f \mid COD)$$

Formel 5: *Berechnung der Wahrscheinlichkeit eines vom COD-Modell erzeugten Alignments unter Berücksichtigung der Unkenntnis des Reading-Frames. Da aber alle Reading-Frames gleich wahrscheinlich sind, ist P(f|COD) = 1/6.*

Für das COD-Modell wird eine spezielle Art von Zustand des pair-Hidden-Markov-Modells benötigt, um „indels", also Insertionen bzw. Deletionen von einem oder mehreren Nukleotiden zu simulieren, die zum Beispiel die durch die zur Gewinnung von Sequenzinformationen genutzte Shotgun-Sequenzierung verursachten Verluste von kurzen Sequenzstücken kompensieren. Weiterhin werden so durch das Alignieren mit blastn hervorgerufene Beschädigungen von Codons repariert, da bei blastn nicht berücksichtigt wird, dass mehrere Nukleotide ein Codon bilden also eigentlich nicht getrennt werden dürfen; blastn kann dies nicht beachten, da das Programm positions-unabhängig arbeitet. Die mit COD bezeichneten Zustände emittieren Codons mit ungleichen Anzahlen von Aminosäuren. In Tabelle 1 sind exemplarisch die Emissionswahrscheinlichkeiten einiger Zustände angegeben.

C(3,3)	$p_{3,3}(a_1a_2a_3, b_1b_2b_3) = p_{COD}(a_1a_2a_3, b_1b_2b_3)$
C(3,2)	$p_{3,2}(a_1a_2a_3, b_1b_2\_) = \sum_{b_3} p_{COD}(a_1a_2a_3, b_1b_2b_3)$
C(3,4)	$p_{3,4}(a_1a_2a_3, b_1b_2b_3b_4) = p_{COD}(a_1a_2a_3, b_1b_2b_3) * p_Y(b_4)$
C(3,0)	$p_{3,0}(a_1a_2a_3, \_\_\_) = \sum_{b_1,b_2,b_3} p_{COD}(a_1a_2a_3, b_1b_2b_3)$

Tabelle 1: *Emissionswahrscheinlichkeiten für die COD-Zustände des COD-Modells. Zu beachten ist, dass es für bestimmte Zahlenverhältnisse der emittierten Nukleotide mehrere Zustände gibt. So gibt es zum Beispiel 3 Zustände C(3,2) und 4 Zustaände C(3,4), je nach dem, an welchen Positionen des Codons die Insertion bzw. Deletion stattfinden soll.*

RNA

Das RNA-Modell repräsentiert die Evolution durch kompensierende Mutationen, die zwar die Sequenz, aber nicht die Sekundärstruktur des Moleküls beeinflussen. Das bedeutet, dass in beiden erzeugten Sequenzen zwischen den identischen Positionen Watson-Crick-Basenpaarungen möglich sind. Um dies zu erreichen, muss ein Mutationsereignis an zwei relativ zueinander definierten Positionen Nukleotide austauschen. Das ist mit Hidden-Markov-Modellen nicht möglich, weswegen hier eine paarweise stochastisch kontextfreie Grammatik verwendet wird. Sie verfügt zum Erzeugen von Sequenzbereichen innerhalb einer Sekundärstruktur über 16*16 Emissionswahrscheinlichkeiten $p_{RNA}(a_L a_R, b_L b_R)$, die jeweils die Wahrscheinlichkeit angeben, dass das Nukleotidpaar ($a_L a_R$) in Sequenz X und das homologe Basenpaar ($b_L b_R$) in Sequenz Y emittiert wird. Ungepaarte Bereiche werden durch die 4*4 Emissionswahrscheinlichkeiten $p_{RNA}(a,b)$ beschrieben, analog zum OTH-Modell. Die Wahrscheinlichkeit für ein Alignment mit der Sekundärstruktur s der RNA berechnet sich dann aus dem Produkt der $p_{RNA}(x_i x_j, y_i y_j)$ für die innerhalb einer Sequenz gepaarten Positionen i und j und der $p_{RNA}(x_k, y_k)$ für die innerhalb einer Sequenz ungepaarten Positionen k. Da die Struktur aber nicht bekannt ist, muss wieder über alle möglichen Strukturen summiert werden. Anders als im Falle der Reading-Frames im COD-Modell sind hier aber nicht alle Strukturen gleich wahrscheinlich. Die Wahrscheinlichkeit einer Struktur für eine gegebene Sequenz wird aber durch einen früher von Rivas und Eddy vorgestellten Algorithmus berechnet. Damit ist es möglich, die Gesamtwahrscheinlichkeit eines von diesem Modell erzeugten Alignments wie in Formel 6 gezeigt zu berechnen.

$$P(\overline{XY} \mid RNA) = \sum_s P(\overline{XY} \mid s, RNA) * P(s \mid RNA)$$

Formel 6: *Berechnung der Wahrscheinlichkeit eines vom RNA-Modell erzeugten Alignments unter Berücksichtigung der Unkenntnis der Sekundärstruktur s. P(s|RNA) wird durch einen Algorithmus von Rivas und Eddy berechnet.*

Die verwendete Grammatik verfügt über die drei Nichtterminale W, W_B und V. V steht für ein Sequenzfragment, dessen Enden gepaart sind, W für eine Teilsequenz, deren Enden eventuell gepaart sind. W_B verhält sich in der Grammatik ebenso wie W, wird aber speziell zum Starten von Multi-Loops eingesetzt. Abbildung 17 zeigt das RNA-Modell mitsamt den in der Grammatik verwendeten Produktionen.

Es gibt zwei verschiedene Arten von Emissionswahrscheinlichkeiten. Für die Emission von gepaarten Nukleotiden in beide Sequenzen werden Wahrscheinlichkeiten vom Typ $p_{RNA}(s_i=(a_L,b_L), s_j=(a_R,b_R))$ verwendet (a jeweils in Sequenz X, b in Sequenz Y). Für ungepaarte Bereiche werden Emissionswahrscheinlichkeiten vom Typ $p_{RNA}(s_i=(a,b))$ genutzt.

Parameterbestimmung

Insgesamt sind 4392 Emissionswahrscheinlichkeiten zu bestimmen. Ideal wäre es, genügend große Trainings-Sets von Sequenzen zu untersuchen, die für ncRNA, Proteine oder keins von beidem kodieren, um diese enorme Anzahl an Parametern zu bestimmen. Leider existieren keine hinreichend großen Sammlungen von Trainingsdaten. Also erfolgte die Parameterbestimmung mit zufällig generierten Sequenzen.
Ausgehend von einer Aminosäure-Substitutionsmatrix (gewählt wurde BLOSUM62) wurden die Wahrscheinlichkeiten der einzelnen die Aminosäuren kodierenden Codons berechnet (p_{COD} aus dem COD-Modell). Summiert man nun für zwei Nukleotide a und b die p_{COD} auf, an denen a und b jeweils an den gleichen Positionen stehen (an der ersten, zweiten bzw. dritten, für alle anderen Nukleotidkombinationen) und dividiert diese Summe durch 3, erhält man die p_{OTH} des OTH-Modells, da es für dieses Modell ja irrelevant ist, an welcher der 3 möglichen Positionen das emittierte Nukleotidpaar (a,b) innerhalb eines Codons steht. Die p_{RNA} des RNA-Modells werden dann ermittelt, indem die p_{OTH} mit den Basenpaarhäufigkeiten kombiniert werden.

Die 48 zu bestimmenden Transitionswahrscheinlichkeiten werden von Hand bestimmt. Dazu werden mit jedem Modell Sequenzen erzeugt und mit realen Vertretern verglichen. Dies ist offensichtlich eine suboptimale Methode, jedoch gibt es bisher dazu keine Alternative.

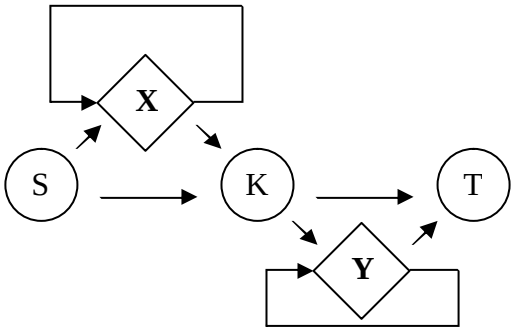

Abbildung 14: *Piktogramm des IID-Modells. S: Startzustand, T: Endzustand, K: Übergangszustand, X: emittiert Nukleotide in Sequenz X, Y: emittiert Nukleotide in Sequenz Y.*

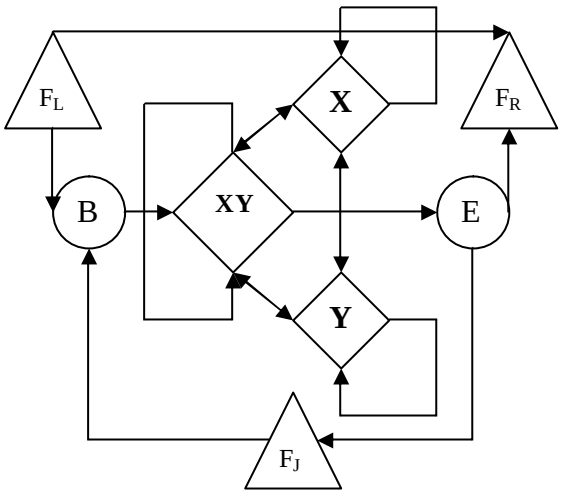

Abbildung 15: *Piktogramm des OTH-Modells. XY emittiert in Sequenz X und Sequenz Y, X nur in Sequenz X, Y nur in Sequenz Y. Die F-Zustände entsprechen jeweils dem IID-Modell und erzeugen unverwandte flankierende Sequenzen.*

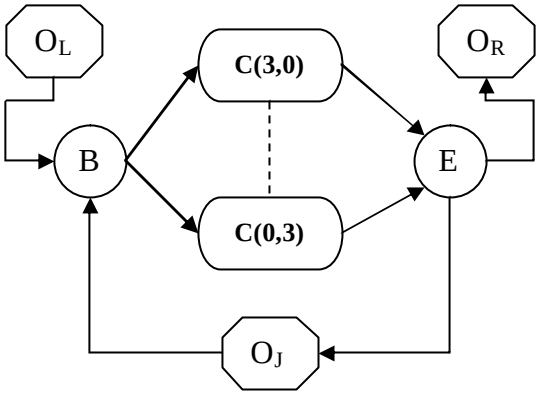

Abbildung 16: *Piktogramm des COD-Modells. Kodierende Bereiche werden eingebettet in nicht kodierende Bereiche generiert, die durch die O-Zustände (entsprechen jeweils dem OTH-Modell) erzeugt werden.*

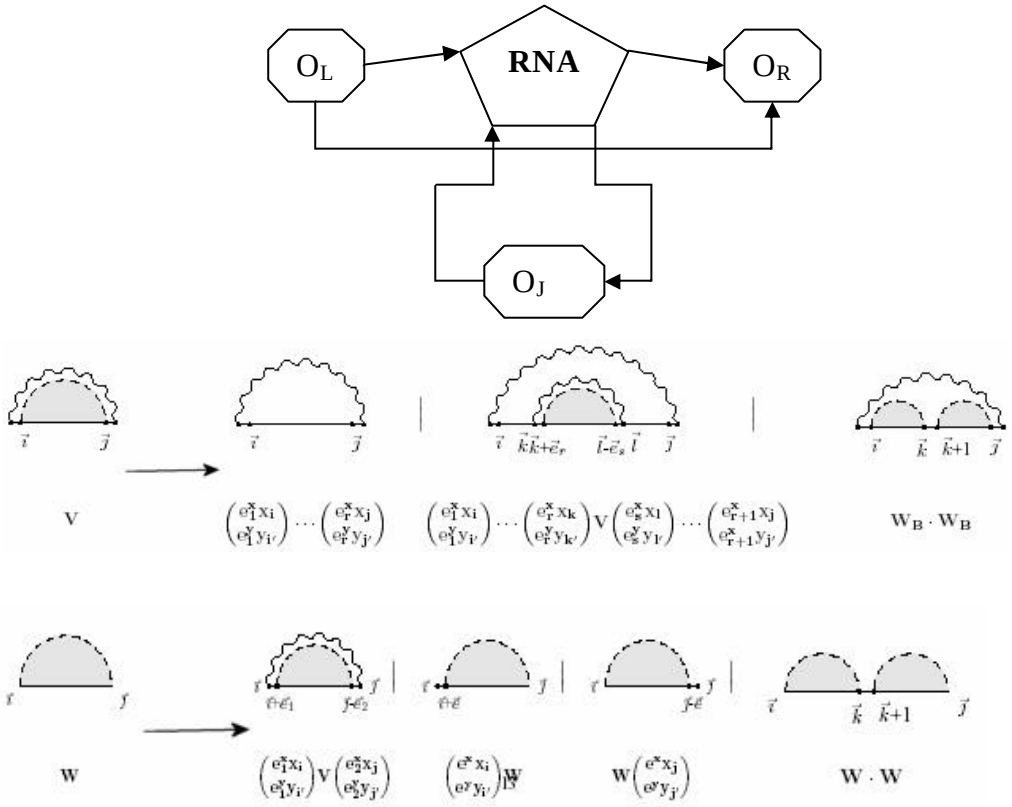

Abbildung 17: *Piktogramm des RNA-Modells. Für ncRNA kodierende Bereiche werden eingebettet in nicht kodierende Bereiche, die durch die O-Zustände (die jeweils dem OTH-Modell entsprechen) erzeugt werden. Die Komponenten des Vektors (e^x, e^y) können beide den Wert 1 oder jeweils nur eine den Wert 0 annehmen. Sie repräsentieren die drei Situationen, dass in Sequenz X und Y, oder in jeweils nur eine der beiden eine Emission erfolgt. Dadurch erklärt sich auch die Indexverschiebung in den Produktionen: wurde der Wert 1 angenommen, wird hier ein nUkleotid emittiert, also muss der nächste Index betrachtet werden; bei 0 entsteht in der Index-Addition wieder der alte Wert. Bei x und y handelt es sich um Nukleotide.*

Die Grammatik startet mit W(i,j). Die Produktionen von W_B verlaufen analaog zu denen von W. V wird erreicht, wenn ein Paar von nukleotiden in beide Sequenzen emittiert wurde, also ein Sekunderstrukturelement begonnen wurde. Die Produktionen von V repräsentieren nacheinander Hairpin, Bulges/Interior Loops bzw. Multibranched Loops.

Alogrithmus

Der Algorithmus erhält als Eingabe das paarweise Alignment der Länge L. Dieses Alignment wird nun unter Verwendung bereits bekannter Algorithmen bewertet. Für die pair-Hidden-Markov-Modelle wird mittels des Viterbi-Algorithmus der Viterbi-Pfad ermittelt, also der wahrscheinlichste Pfad durch die Zustände des Modells mitsamt seiner Wahrscheinlichkeit. Für die stochastisch kontextfreie Grammatik des RNA-Modells findet eine Kombination aus CYK- und Inside-Algorithmus Verwendung. Der CYK-Algorithmus löst das Wortproblem kontextfreier Sprachen (hier: ob das Wort durch die Grammatik erzeugt werden kann), und der Inside-Alogirthmus berechnet die Satzwahrscheinlichkeit, also in unserem Fall die Wahrscheinlichkeit, dass das gegebene Alignment von diesem Modell erzeugt wurde. Somit lässt sich für ein gegebenes Input-Alignment für alle drei Modelle die Wahrscheinlichkeit bestimmen, das Alignment erzeugt zu haben, oder anders gesagt der Wahrscheinlichkeit des Alignments bedingt auf das jeweilige Modell. Dies entspricht der Wahrscheinlichkeit, dass das Alignment dem durch das Modell repräsentierten Sequenztyp zuzuordnen ist. Unter der Annahme, dass a priori alle drei Modelle gleich wahrscheinlich sind, lässt sich nun die Bays'sche Posteriori-Wahrscheinlichkeit berechnen, die die Wahrscheinlichkeit angibt, dass das jeweilige Modell (bzw. die korrespondierende Gen-Kategorie) dem Alignment zuzuordnen ist, oder in bedingter Wahrscheinlichkeit ausgedrückt: die Wahrscheinlichkeit des jeweiligen Modells bedingt auf das gegebene Alignment. Formel 7 zeigt die entsprechende Berechnung.

A $$P(Model_i) = \frac{1}{3}$$

B $$P(\overline{XY}) = \sum_{i \in \{OTH, COD, RNA\}} P(\overline{XY} \mid Model_i) * P(Model_i)$$

C $$P(Model_i \mid \overline{XY}) = \frac{P(\overline{XY} \mid Model_i) * P(Model_i)}{P(\overline{\overline{XY}})}$$

Formel 7: *Berechnung der Bays'schen Posteriori-Wahrscheinlichkeit (C), die angibt, wie wahrscheinlich es ist, dass das betrachtete Modell dem beobachteten Alignment zuzuordnen ist. a priori werden alle Modelle als gleich wahrscheinlich angenommen (A). B gibt an, wie die Wahrscheinlichkeit des Alignments berechnet wird.*

Zur Beurteilung der Güte der Ergebnisse wird unter der Annahme, dass alle eingegebenen Sequenzalignments dem OTH-Modell entspringen je ein log-odds-Score für das COD- bzw. das RNA-Modell berechnet, der angibt, wie nahe die Erwartung am beobachteten Ergebnis liegt, und diese zwei Werte in einem „Phase Diagram" gegeneinander aufgetragen. Formel 8 zeigt die entsprechende Berechnung.

$$(x, y) = \left(ld \frac{P(COD \mid \overline{XY})}{P(OTH \mid \overline{XY})}, ld \frac{P(RNA \mid \overline{XY})}{P(OTH \mid \overline{XY})} \right)$$

Formel 8: *Berechnung der logg-odds-Scores für das COD- bzw. das RNA-Modell unter der Annahme, dass alle analysierten Sequenzalignments weder für RNA noch für Proteine kodieren, also dem OTH-Modell zuzuordnen sind.*

Für die Zugehörigkeit eines Alignments zu einem Sequenztyp lassen sich feste Beziehungen zwischen diesen beiden Scores angeben: gilt y>x und y>0, so ist das Alignment codierend für ncRNA; ist x>y und x>0, so ist es kodierend für Proteine; sind x und y beide kleiner 0, so handelt es sich um nicht kodierende Bereiche. So lassen sich die Punkt des Phase Diagrams nach ihrer Position klassifizieren.

Tests

Zunächst wurde das Verfahren mit Sequenzen getestet, die von den drei Modellen erzeugt wurden. Sinn dieses Tests war es, herauszufinden, wie leistungsfähig der Algorithmus ist bzw. wie viele Alignments von ihm falsch klassifiziert werden. Solche Aussagen können natürlich nur getroffen werden, wenn die Klasse des eingegebenen Alignments bereits im Voraus bekannt ist. Mit jedem Modell wurden dazu 1000 jeweils 200 Nukleotide lange Alignments erzeugt. Um zu zeigen, dass die vorgenommene Klassifizierung nur auf dem simulierten (bzw. analysierten) Mutationsmuster beruht, wurden die Alignments für einen zweiten Versuch umgestaltet, indem die jeweils gegeneinander alignierten Positionen spaltenweise vertauscht wurden. Die prozentuale Sequenzidentität und die Anzahl der Gaps blieben so konstant, doch die Codon-Struktur, die für kodierte Aminosäuren wichtig ist, und die für die Sekundärstruktur nötigen Bindungsbeziehungen zwischen bestimmten Positionen wurden so zerstört. Die Ergebnisse wurden in den in Abbildung 18 gezeigten Phase-Diagramms eingetragen.
Vor dem Shuffling der Alignments ist die Klassifizierung nahezu immer korrekt und nur wenige Ausreißer sind zu beobachten. Nach dem Shuffling werden die Alignments mehr oder weniger eindeutig dem OTH-Modell zugeordnet. Dies war auch nicht anders zu erwarten, da die für die Einordnung in andere Klassen nötigen Mutationsmuster innerhalb der Sequenzen zerstört wurden.

Der nächste Test erfolgte mit virtuellen Genomen. Sinn war es, die Tendenz zu „false Positives“ zu beobachten, wenn Sequenzbereiche der Modelle OTH und COD dominieren. Zu diesem Zweck wurden zwei Pseudobakteriengenome generiert, die keine ncRNA-Gene enthalten. Jedes hat eine Länge von 2000 Nukleotiden und besteht zu 90% aus vom COD-Modell erzeugten Sequenzstücken, die restlichen 10% entspringen dem OTH-Modell. Es wurden drei Paare dieser Pseudobakteriengenome generiert, in denen jeweils der GC-Gehalt variiert wurde; er lag bei

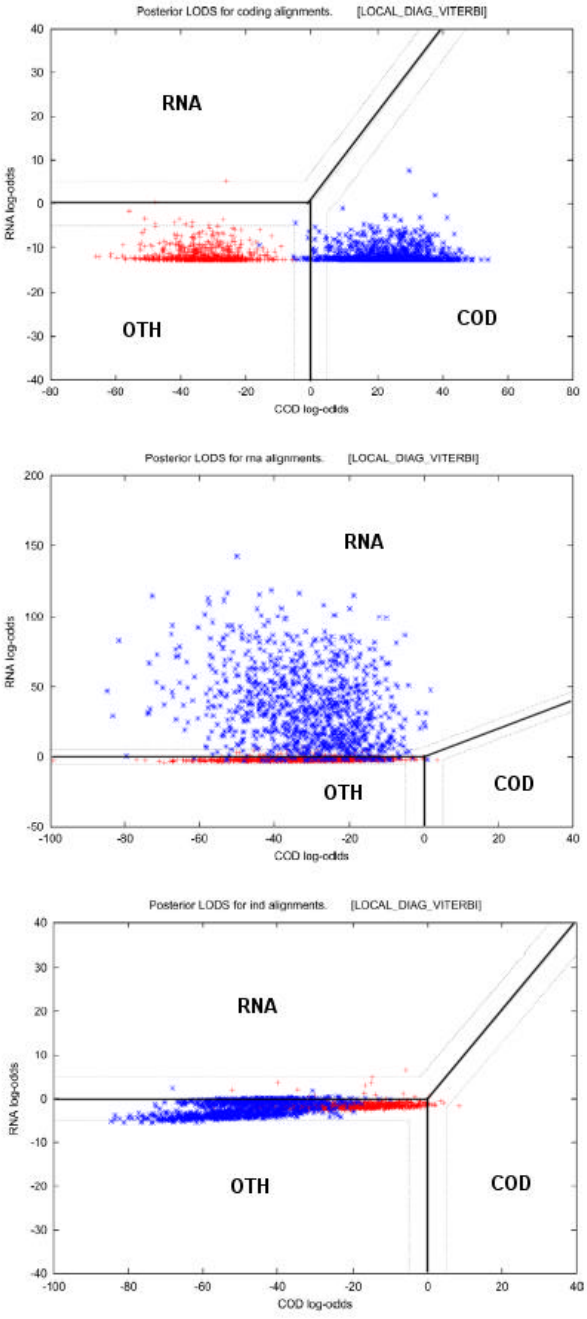

Abbildung 18: *Phase Diagrams des Tests mit den Genomen von Escherichia coli und Salmonella typhi. Die blauen Markierungen repräsentieren klassifizierte Alignments vor dem Shuffling, die roten Markierungen repräsentieren klassifizierte Alignments nach dem Shuffling.*

38.9, 47.25 bzw. 57.7 Prozent. Die Anzahl der false Positives stieg mit dem GC-Gehalt von 1 über 14 bis zu 21. Der Genomsatz mit dem höchsten GC-Gehalt wurde nochmals analysiert, nachdem die Parameter der Modelle hinsichtlich dieses Kriteriums neu eingestellt wurden. Hier ergab sich lediglich eine fälschlicher Weise positive Erkennung. Die mit steigendem GC-Gehalt sinkende Sensitivität kann also durch eine Veränderung des Parametersatzes ausgeglichen werden.

Der dritte Test, diesmal unter realen Bedingungen, erfolgte mit den Genomen von *Escherichia coli* und *Salmonella typhi*, zwei verwandten Arten von Enterobakterien. Von *Escherichia coli* ist bekannt, dass das Genom 115 ncRNA-Gene sowie 4290 für Proteine kodierende Gene enthält. Entsprechend wurden die Genombereiche in 3 Sammlungen von Sequenzen eingeteilt, getrennt nach der Kodierung für ncRNA, für Proteine bzw. nichts. Jede der drei Sammlungen wurde mittels blastn gegen das komplette Genom von Salmonella typhi aligniert und anschließend durch das Verfahren bewertet.
Von den 115 bekannten ncRNA-Genen wurden lediglich 33 in Alignments mit einem e-Wert unter 0,01 und einer Länge von mehr als 50 Nukleotiden, also in signifikanten Alignments, wiedergefunden. Von diesen 33 wurden aber alle richtig als ncRNA-Gene klassifiziert. Von den 4290 für Proteine kodierenden Genen waren 3181 in signifikanten Alignments enthalten, von denen 2876 richtig zugeordnet wurden.
Insofern die zu untersuchenden Sequenzen in signifikanten Alignments enthalten waren und so analysiert werden konnten, wurde also ein Großteil von ihnen korrekt klassifiziert. Da das Verfahren entwickelt wurde, um ncRNA-Gene zu erkennen, ist es nicht verwunderlich, dass die Anzahl der Fehlklassifikationen im Falle der für Proteine kodierenden Gene zunimmt.

Fazit

Elena Rivas und Sean Eddy haben in ihrer Veröffentlichung „Noncoding RNA gene detection using comparative sequence analysis” ein Verfahren vorgestellt, das es ermöglicht, in einer Laufzeit von $O(n^3)$ und einem Speicherbedarf von $O(n^2)$ ncRNA-Gene zu identifizieren. Auftretende Fehler lassen sich dadurch erklären, dass die Parameterbestimmung an eine Aminosäuresubstitutionsmatrix mit einer festen evolutionären Distanz gebunden ist, die nicht mit der Distanz zwischen allen untersuchten Sequenzen übereinstimmen kann. Weiterhin beachtet das zum Alignieren verwendete blastn aufgrund seines poitionsunabhängigen Vorgehens nicht, dass Nukleotide eines für eine Aminosäure kodierenden Codons nicht getrennt werden dürfen, also entweder das gesamt Codon zum Alignment

gehört oder nicht. So kann es sein, dass die Alignments weniger Nukleotie umfassen als sie eigentlich sollten. Da das Verfahren nach stark konservierten Sekundärstrukturen sucht, kann es eventuell zu Fehlklassifikationen von mRNA mit stark ausgebildeter Sekundärstruktur als ncRNA kommen.
Der Algorithmus ist implementiert in QRNA, einem Prototyp eines ncRNA-Gene findenen Programms. Weiterhin ist es möglich, auf die gleiche Art und Weise für Proteine kodierende Gene zu identifizieren. Dies bedarf aber noch weiterer Modifikationen des Modells, um die Detektion verlässlicher zu machen.

Quellen

[1] Ye Ding, Charles Lawrence: **A statistical sampling algorithm for RNA secondary structure prediction**; *Nucleic Acids Research* 31(24), S.7280-7301.

[2] Elena Rivas, Sean Eddy: **Noncoding RNA gene detection using comparative sequence analysis**; *BMC Bioinformatics* 2(8).